Studies of Extensions of HRM-SDT for Constructed Responses

Xiaoliang Zhou

Submitted in partial fulfillment of
the requirement for the degree of Doctor of Philosophy
under the Executive Committee of
the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2019

ABSTRACT

Studies of Extensions of HRM-SDT for Constructed Responses

Xiaoliang Zhou

This research examines an ordered perception rater model, an extension of the equal perception signal detection theory (SDT) latent class rater model. The expectation-maximization algorithm and the Newton-Raphson algorithm are used to estimate parameters. Four simulation studies are conducted to answer three research questions.

Simulation studies 1 and 2 fit correct models to the data. Simulation study 1 generates one hundred data sets from the equal perception rater model, both with fully-crossed design and BIB design, and both without and with rater effects, and fits the equal perception model. Parameter recovery is excellent for fully-crossed design and reasonable for BIB design, and all rater effects are detected. Simulation study 2 generates one hundred simulated data sets from the ordered perception model, both with fully-crossed design and BIB design, and both without and with rater effects, and fits the ordered perception rater model. Although parameter recovery is biased for some parameters in the BIB design, all rater effects are recovered.

Simulation studies 3 and 4 fit wrong models to the data. Simulation study 3 fits equal perception models to the fully-crossed and BIB ordered perception data sets generated in simulation study 2. All rater effects are revealed, although rater effects are distorted to some extent in the BIB design. Simulation study 4 fits ordered perception models to the fully-crossed and BIB equal perception data sets generated in study 1. All rater effects are recovered.

Using essay scores from a large-scale language test, an empirical study is conducted. Both the equal and the ordered perception models are fitted. Information criteria favor the equal perception model.

Contents

List of Tables

List of Figures

Acknowledgments

I would like to give my foremost thanks to my adviser, Dr. Lawrence T. DeCarlo. Without your consistent support, encouragement, and comments, I would have never been able to complete this dissertation. It is a great pleasure to listen to your smart comments on the current development of psychometrics and statistics, your views on life and human nature, and your advice on how to prepare for presentations and job interviews. It is of tremendous fun to attend your seminars where it is impossible to feel tired or drowsy once you start to talk. Also, I have learned essential measurement concepts from your courses of psychological measurement, latent structure analysis, and multilevel and longitudinal data analysis. You have the magic of keeping students laughing now and then with your humorous remarks so we understood the most important concepts with a light heart and little burden.

Also, I would like to extend my sincere thanks to Dr. Mathew S. Johnson, Dr. Arron M. Pallas, Dr. Charles Lang, and Dr. Shaw-Hwa Lo, the members of my thesis committee, for your insightful comments and constructive suggestions. Dr. Johnson's deep understanding of statistics made it enjoyable to learn complex ideas. Dr. Pallas's decades of studies on education introduced me to a fantastic world of sociology of schools. Dr. Lang showed to me how to apply data mining skills to real world education data. Dr. Lo gave the modern statistical techniques course which I found to be the most thorough and deep. Thank you, Lo, for inviting me to your thanksgiving parties where I knew many talented researchers and scholars.

I would also like to say thanks to the fabulous professors at TC and QMSS who helped me to grow academically by showing me how to do research or teach. Thanks first go to professors in the Department of Human Development, Dr. James E. Corter, Dr. Young-Sun Lee,

Dr. Bryan S. Keller, and Dr. Laura E. Tipton. I have learned from all of you basic to advanced statistics or psychometrics. Thank you Corter for meeting me regularly and guiding me to do a project on CDM. I would say special thanks to Dr. Alex J. Bowers in the educational leadership program who provided me with fellowship and graduate assistantship from 2016 to 2018, teaching me hands-on skills of doing statistical analysis and visualizing education data, introducing me to education and big data conferences, and greatly relieving my financial burden of study at TC. Thanks also go to Dr. Mun C. Tsang who gave me a graduate assistantship to do two studies on Children's museum in 2016. Finally, I would like to thank Dr. Parrott, Dr. Eirich, and Dr. Goodrich at the QMSS program who gave me abundant freedom of holding office hours or attending classes, especially after I underwent two surgeries.

I would like to thank my friends and classmates I came to know while studying at TC. I enjoyed the numerous discussions with Nayeon Yoo, Rui Lu, Jiaxi Yang, Xiang Liu, Zhuangzhuang Han, Lu Han, Yihan Zhao, Xinyu Ni, Sen Zhang, Tianyang Zhang, Yilin Pan, and many others. You lightened and enlightened my life at TC.

Finally, I would like to thank my parents, Aimei and Fawen, for your dedicated love, for the examples you set for me, always to be diligent, simple, grateful, always to strive for the best. I would like to thank Cuiyan my sister who took care of my parents while I was studying at Columbia University. A special thank-you goes to my love Le. Without Le's companionship and care for me, I might have been tortured more by depression and anxiety and would have been unable to complete this work at this time.

To Aimei and Fawen

Chapter I

Introduction

Testing with constructed response (CR) items has a very long history and is still popular today. The traditional Chinese *keju* imperial exam lasted for over 1,300 years, where examinees wrote several essays and were evaluated by one or two official experts. In this period, essays were the only form of assessment. Today multiple choice (MC) items constitute a large proportion of tests, however testing with CR items is still acknowledged as a necessary way of evaluating performance in many modern tests such as the SAT (College Board), ACT (ACT), GRE (Educational Testing Service), and Advanced Placement Program (AP; College Board). Besides essays, other types of CR items are open-ended questions and ratings of art works.

Testing institutions include CR items in tests owing to three advantages that they have over MC items. First, the time an examinee takes to complete a CR item is 16 times longer than the time to complete an MC item (Lukhele, Thissen, & Wainer, 1994), which means that the examinee has sufficient time to generate in-depth information (Pollock, Rock & Jenkins, 1992; Rodriguez, 2002). Therefore, stakeholders such as students, parents, teachers, and principals can have a better understanding of students' academic achievements. Second, it is difficult for examinees to guess on a CR item, and even if they write prepared answers, it is easy for trained raters to detect such answers. This means that CR items tend to generate more valid information than MC items on the constructs examiners wish to obtain information about. Statistically, no guessing parameters are needed releasing more degrees of freedom. Third, CR items can measure information on test takers of extremely high or low abilities (Ercikan et al., 1998). As

long as an examinee gives a response to a CR item, raters will always have some information to give a rating score.

Despite these advantages of CR items over MC items, two characteristics of CR items make it more difficult to model scores from CR items as compared to those from MC items. First, whereas MC items can be automatically and objectively scored as right or wrong, coded as 1 or 0, the quality of CR items needs to be judged by raters using complex cognitive processes. Second, the cognitive scoring process involves raters evaluating the CR, which means that raters may have biases or inconsistently score the same CR, and they may do either thing in different degrees.

A typical model to estimate essay quality is the latent trait model, an IRT-based model (Linacre, 1989; Muraki, 1992). For example, the FACETS model (Linacre, 1989) is one of the earliest efforts to model rater effects in addition to item and examinee characteristics. However, this model has a major problem in its way of accumulating information (Casabianca, Junker, & Patz, 2012), in that perfect estimation of examinee proficiency is achievable by increasing the number of ratings while holding the number of items constant (DeCarlo, Kim, & Johnson, 2011; Mariano, 2002). This way of treating the response data is problematic since this assumes independence among these ratings but actually the multiple ratings of the same item are dependent (Casabianca et al., 2012). Raters are also correlated across essays since raters are rating responses to the same prompt and have received the same training for scoring.

There have been attempts to address rating dependence by applying a hierarchical structure and combining IRT and generalizability theory (GT; Brennan, 1992, 2001). For example, researchers have proposed an IRT model for multiple raters (IRT-MMR) that considers each observed item-examinee combination as representing a latent continuous quality of the

examinee who responds to an item (Verhelst & Verstralen, 2001). Other researchers have developed alternative forms of multiple ratings IRT models using "rater bundles" (Wilson & Hoskens, 2001). Yet other researchers have attempted other forms of utilizing the GT model (Bock, Brennan, & Muraki, 2002; Briggs & Wilson, 2007). Still other studies have made attempts to extend the FACETS model (see Hung & Wang, 2012; Mariano & Junker, 2007; Muckle & Karabatsos, 2009; Wang & Liu, 2007; Wang & Wilson, 2005).

The present study examines essay quality within a hierarchical rater model (HRM; DeCarlo, Kim, & Johnson, 2011; Patz, 1996; Patz, Junker, Johnson, & Mariano, 2002). Major advantages of HRM lie with its ability to appropriately model dependence among multiple ratings of the same essay and it also solves the information accumulation problem that exists in many IRT-based models.

HRM has been extended in two directions, one by adding covariates of the rating process such as rater status (e.g., human vs. machine, see Casabianca et al., 2012; Wang, 2012) into the model, the other by incorporating a discrete latent variable between the examinee's discrete latent quality of the CR item and the observed rater scores (DeCarlo, 2002, 2005; DeCarlo et al., 2011). The first direction of HRM extension (HRMC) comes from Junker and his colleagues (Casabianca & Junker, 2013; Mariano & Junker, 2007; Patz et al., 2002). This effort makes it possible to incorporate factors into HRM that may influence rater effects such as bias and variability. The idea of including covariates in latent class models has a history of over three decades (Bandeen-Roche, Miglioretti, Zeger, & Rathouz, 1997; Clogg & Goodman, 1984; Dayton & Macready, 1988a; Dayton & Macready, 1988b; Formann, 1992; Huang & Bandeen-Roche, 2004; Kamakura, Wedel, & Agrawal, 1994; Melton, Liang, & Pulver, 1994; Yamaguchi, 2000).

Another approach to HRM comes from DeCarlo and his colleagues who borrow ideas from signal detection theory (SDT; DeCarlo, 2002, 2005; DeCarlo et al., 2011), leading to an HRM-SDT model. SDT is a cognitive theory of perception that is consistent with various differences among raters. Specifically, a latent perception variable is assumed to mediate between the observed rating scores and the ideal category of the CR item. In addition, a rating is arrived at by using decision criteria. By varying the location of the decision variable, researchers can model rater effects of bias, variability, and tendency to favor certain score categories (DeCarlo, 2008b; Wolfe & McVay, 2012). The inclusion of the perception and decision variables offers advantages over the original HRM in terms of being able to model rater effects and rater accuracy.

The present study adopts and extends the HRM-SDT model by adding some flexibility. The main manipulation is to allow the distance parameter to vary. Allowing flexibility in the distance parameter may be worth trying since there is no necessary reason, apart from parsimony, to fix this parameter. We examine parameter recovery for this model and see how it affects the simplified equal perception model.

Models tend to have better performance as the number of parameters increase and so relative fit measures recognize this by penalizing for the number of parameters. Researchers commonly use measures such as Akaike information criterion (AIC; Akaike, 1973), Bayesian information criterion (BIC; Schwarz, 1978), or deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Van der Linde, 2002) to assess relative fit. For these measures, smaller values mean better model fit.

The present study conducts both simulations and real data analyses to explore the ordered perception model. In the first and fourth simulation studies, the same equal perception data are

fitted with both the equal and ordered perception models and results are compared. In the second and third simulation studies, the same ordered perception data are fitted with both the ordered and equal perception models and results are compared. Analyses of real data are carried out to see whether the ordered perception model outperforms the equal perception model. Thus, the present study addresses the following research questions:

1. To what extent can model parameters be recovered for the ordered perception model?
2. To what extent will fitting wrong models affect parameter recovery?
3. How does model fit compare for the ordered and equal perception models?

Chapter 2 reviews previous models used to model CR items, including the FACETS model, the HRM, the HRM-SDT, the HRMC, and the LC-SDTC, and introduces the ordered perception model. This chapter also reviews diagnostic measures for evaluating model performance, such as AIC, BIC, and DIC, as well as estimation methods such as maximum likelihood (ML), marginal maximum likelihood (MML), and Markov chain Monte Carlo (MCMC). Chapter 3 outlines methods of assessing the ordered perception model, including simulation and empirical studies. Chapter 4 reports results of the simulation and empirical studies. Chapter 5 summarizes findings of the current study and discusses implications and limitations.

Chapter 2

Literature Review

Unlike MC items, which are objectively scored as right or wrong, usually coded as 1 or 0, CR items tend to be scored by different raters, or the same rater may rate more than one CR item. Under this circumstance, it is inevitable that rater effects will affect the scoring of CR items. Common rater effects include rater severity, rater centrality/extremity, restriction of rating range, and halo effect (Engelhard Jr, 1994, 1996). To incorporate these rater effects into appropriate models, researchers have put forward a plethora of rater models. One of the earliest such models is the FACETS model.

2.1 The FACETS Model

The FACETS model (Linacre, 1989) incorporates rater severity into the modeling process. It is a Linear Logistic Test Model (LLTM) (Fischer, 1973, 1983) based on item response theory (IRT). It treats raters, items, and examinees effects on the logit scale, in the same way as does an ANOVA approach (Patz et al., 2002). Through parameters in its function, the FACETS model recognizes that score variability is derived from sources or "facets" such as items, examinees, and raters (Linacre, 1989). The main rater effect incorporated into the model is rater severity,

$$\log\left[\frac{P(Y_{ijr}=k+1|\theta_i)}{P(Y_{ijr}=k|\theta_i)}\right]=\theta_i-\beta_j-\gamma_{jk}-\phi_r \qquad (1)$$

where

$P(Y_{ijr}=k+1|\theta_i)$ = probability of rater *r* giving examinee *i* score *k+1* on item *j*,

$P(Y_{ijr} = k|\theta_i)$ = probability of rater *r* giving examinee *i* score *k* on item *j*,

θ_i = latent proficiency of examinee *i*,

β_j = difficulty for item *j*,

γ_{jk} = step parameter for item *j*,

ϕ_r = bias or severity for rater *r*.

So far, there have been a host of studies using the FACETS model and its extensions to explore various rater effects, such as rater severity and differential rater functioning (Engelhard Jr, 1994, 1996; Jin & Wang, 2017; Myford & Wolfe, 2003, 2004; Wesolowski, Wind, & Engelhard Jr, 2015, 2016; Wu & Tan, 2016). Differential rater functioning occurs when raters demonstrate differing degrees of severity in rating students from different subgroups, such as gender subgroups or ethnic subgroups.

Popular as it is, the FACETS model tends to have several disadvantages. First, it may have made inappropriate assumptions that raters have an equal ability to discriminate among categories of CR items. However, raters may vary on discrimination, since when scoring CR items, raters may be influenced by factors such as expertise in scoring CR items and emotional fluctuations and other factors such as duration of training and environmental stimuluses.

Second, the FACETS model can miss information about rater effects. The FACETS model only accounts for rater severity/leniency, whereas other effects, such as avoiding end categories (centrality), are often found in practice. If the FACETS model covers only one rater effect, then this model will miss other rater effects.

Third, the FACETS model ignores the fact that multiple raters, when rating the same item, are not independent. Owing to the improper assumption of rater independence, the item difficulty

estimates shrink toward zero, a problem especially serious for extreme items, and the estimates for the variance of examinee proficiency distribution are also underestimated (Patz et al., 2002).

Fourth, a fundamental flaw in the FACETS model is that increasing the number of raters will indefinitely increase the precision of estimating examinee proficiency, even when there is only one CR item in the test (DeCarlo et al., 2011; Patz et al., 2002). It has also been argued that increasing the number of ratings can only increase the precision of estimating the latent quality of the CR item, and not the precision of examinee proficiency (DeCarlo et al., 2011).

2.2 The Hierarchical Rater Model (HRM)

To tackle issues arising from assumptions of rater independence, and to effectively model rater effects, researchers (Patz, 1996; Patz et al., 2002) modified the traditional IRT model and proposed the HRM model. As shown by Equation 2, the first level of HRM corresponds to the process of rater *r* scoring CR *j*, linking rater scoring Y_{ijr} to latent examinee proficiency η_{ij} associated with specific constructed item *j*. The second level corresponds to the process of examinee *i* responding to CR item *j*, connecting latent proficiency η_{ij} with examinee ability θ_i.

$$\begin{aligned} &\theta_i \sim i.i.d.\ (\mu, \sigma^2), \text{ for examinee } i \\ &\eta_{ij} \sim \text{IRT model, for examinee } i \text{ and item } j \\ &Y_{ijr} \sim \text{SDT model, for examinee } i, \text{ item } j, \text{ and rater } r. \end{aligned} \tag{2}$$

Patz et al.'s (2002) version of the HRM model is shown in Figure 2.1. As shown in the figure, on the second level the latent examinee proficiency θ is reflected through the latent quality of CR items η where the nonlinear relationship is realized through parameters *a* and *b*. On the first level, the latent quality of CR items η is in turn evaluated and given scores Y by multiple raters where the nonlinear relationship is modeled through parameters ϕ and Ψ.

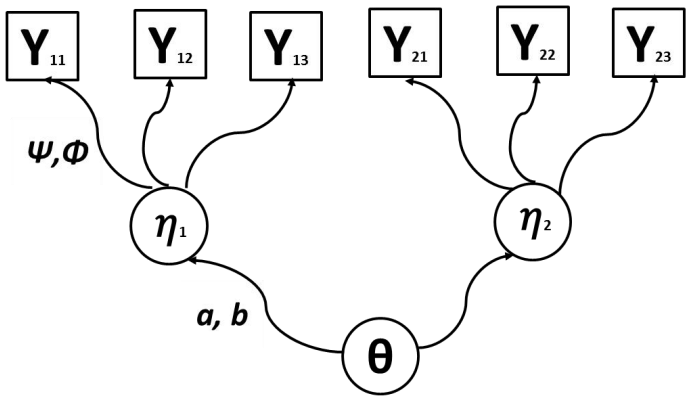


Figure 2.1. Representation of HRM (Patz et al., 2002)

First level: Simple SDT model

In the HRM designed for polytomous items (Mariano & Junker, 2007; Patz et al., 2002), the first level employs a simple signal detection theory model where a rater assesses an examinee' CR and assigns to it a score reflecting its quality, as defined by a scoring rubric.

The relationship between the ideal ratings and the observed ratings in the first level of the HRM can be represented in a matrix (Table 2.1). Here, the HRM uses a discrete latent model with four categories to show the quality of the ratings for the CR. As mentioned above, $p_{\eta kr}$ means the probability of rater *r* giving score *k* given ideal rating η. Ideally, the entries on the diagonal of the matrix should be close to 1 and the off-diagonal entries should be close to 0, which means that the rater can accurately capture the ideal latent category of the CR through the ratings she gives.

Table 2.1. Rating Probabilities for the First-level Signal Detection Process Modeled in the HRM

Ideal Rating (η_{ij})	Observed Rating (k)			
	0	1	2	3
0	p_{11r}	p_{12r}	p_{13r}	p_{14r}
1	p_{21r}	p_{22r}	p_{23r}	p_{24r}
2	p_{31r}	p_{32r}	p_{33r}	p_{34r}
3	p_{41r}	p_{42r}	p_{43r}	p_{44r}

The HRM uses a discrete unimodal distribution to model each row of the matrix, or the probability of the observed rating, $P[Y_{ijr} = k|\eta_{ij} = \eta]$. The mode of the distribution represents the bias ϕ_r of the rater r, and the spread of the distribution the variability Ψ_r of the rater *r*. This probability can be modeled with a normal distribution, with $P[Y_{ijr} = k|\eta_{ij} = \eta] \sim N(\eta + \phi, \Psi_r)$. So, the first level of HRM can be written as

$$p_{\eta kr} = P[Y_{ijr} = k|\eta_{ij} = \eta] \propto \exp\{-\frac{1}{2\Psi_r^2}[k - (\eta + \phi_r)]^2\} \quad (3)$$

where

$p_{\eta kr}$ = probability of rater *r* giving score *k* given ideal rating η,

Ψ_r = variability for rater *r*,

ϕ_r = bias or severity for rater *r*.

The inverse of Ψ_r^2 measures the rater precision. The higher this inverse, the more reliable rater *r* is. The parameter ϕ_r indicates the bias of rater *r.* A value of zero means that rater *r* has no bias. Positive values mean that rater *r* is likely to be lenient and give scores higher than latent class η, whereas negative values mean that rater *r* is likely to be strict and give scores lower than latent class η.

Though HRM has appropriately relaxed the restricted assumption of rater independence made in the FACETS model, it is not free from estimation problems. First, raters with small values of Ψ tend not to have good estimates of ϕ (Patz et al., 2002). For small values of Ψ, the likelihood function of ϕ is approximately uniform over (-0.5, 0.5), which means that it is difficult for the parameter estimation algorithm to select a value for ϕ within this interval (Patz et al., 2002). Beyond this interval, the likelihood drops to nearly zero because the probability of scores in response categories other than the true category is close to zero (DeCarlo et al., 2011).

Second and more importantly, while modeling rater severity/leniency, HRM ignores such rater effects as restriction of rating range or central tendency, which commonly appear in real world data (DeCarlo et al., 2011).

Second level: IRT model

For an item with polytomous response categories, the second level of HRM uses polytomous IRT models, such as the partial credit model (PCM) (Masters, 1982) or the generalized partial credit model (GPCM) (Muraki, 1992), to link the examinee ability to latent categories of this item. Both models are generalized linear models using adjacent category logits (Agresti, 2013; Dobson & Barnett, 2008). The GPCM represents the ratio of two adjacent latent categories on the logit scale and can be written as

$$\log\left[\frac{P(\eta_j = \eta + 1|\theta)}{P(\eta_j = \eta|\theta)}\right] = a_j\theta - b_{jm} \tag{4}$$

where

$P(\eta_j = \eta + 1|\theta)$ = probability of examinee being in latent category $\eta + 1$,

$P(\eta_j = \eta|\theta)$ = probability of examinee being in latent category η,

η_j = latent category for item j, at 0, 1, …, M–1,

θ = examinee proficiency, assumed as $N(0, 1)$,

a_j = item discrimination for item j,

b_{jm} = item step for item j ($m = \eta - 1$).

Modifying Equation 4 produces other polytomous IRT models (DeCarlo, Kim, & Johnson, 2011). For example, restricting a_j to be equal for all items gives rise to the PCM, and using cumulative probabilities leads to the graded response model (GRM) (Samejima, 1969).

2.3 The Hierarchical Rater Model with a Latent Class Signal Detection Theory (HRM-SDT)

The problems of HRM such as parameter estimation difficulties and the inability to estimate rater effects other than severity have been addressed (DeCarlo et al., 2011) through incorporating a latent class SDT into the first level of HRM (DeCarlo, 2002, 2005, 2008a). The usefulness of applying SDT (Green & Swets, 1988) to understanding the psychological process of scoring CR items has been shown in previous studies (DeCarlo, 2002, 2005).

Incorporating latent class SDT into the first level of HRM

SDT involves two processes, namely, the perception of an item's quality and the use of criteria to score this item, associated with parameters *d* and *c* respectively. The use of these two parameters is illustrated in Figure 2.2. It is assumed that the quality of response to the item is a latent continuous variable η, and the perceptions Ψ of the response's quality are based on a location-family probability distribution, such as logistic or normal (DeCarlo, 1998). For a latent variable η with four latent categories, the location-family distribution has four locations, each for one latent category. In the task of detecting an item from latent category *m*, the rater's perceptions Ψ are based on the *m*th probability distribution. Parameters *d* and *c* correspond to rater precision and response criteria respectively.

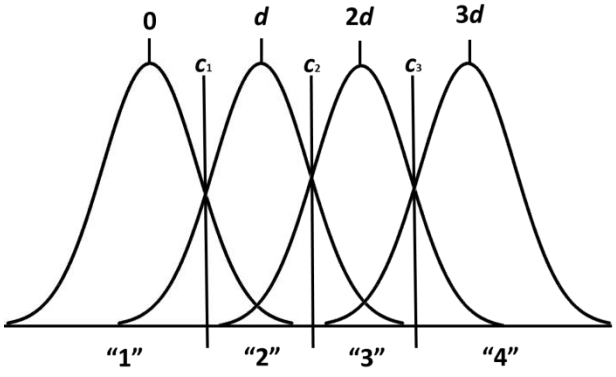


Figure 2.2. Representation of SDT (DeCarlo, 2002, 2005)

Parameter *d*, the distance between two locations of perceptual distributions, represents the rater's ability to distinguish the latent categories of an item, hence an indicator of rater precision. Although previous research has shown that an equal perception assumption is useful in parameter estimation (DeCarlo, 2002, 2005), it is also possible to assume that the distances between the perceptual distributions vary since it is unknown whether data with true ordered perceptions will be better modeled with ordered perception models or equal perception models. However, in the HRM-SDT model, equal perception is assumed. The present study examines effects of ordered perceptions.

On the other hand, parameter *c* provides reference points that "divide the decision space into the four response categories" (DeCarlo et al., 2011, p. 338), corresponding to the four scores of 1 to 4. If the rater perceives that the quality of an item is between the 2nd and 3rd criteria, for example, then the rater gives the examinee a score of 3. In Figure 2.2, the location of *c* is located at the intersection position of distributions which is "optimal" in the sense that it will maximize proportion correct in certain circumstances (DeCarlo, 2008a; DeCarlo et al., 2011).

Based on the discussion above and the information in Figure 2.2, the latent class SDT model for the cumulative probability of rater *j* giving score *k* to an item can be written as

$$p(Y_j \leq k \mid \eta = \eta) = F(c_{jk} - d_j \eta) \tag{5}$$

where

Y_j = score given by rater *j*, at 0, 1, …, *K*–1,

η = latent ordinal category for an item,

F = cumulative location-family distribution function,

c_{jk} = response criteria for category *k* and rater *j*,

d_j = precision for rater *j*.

Note that there are two standard assumptions in ordinal response models. First, c_{jk}'s are strictly ordered, with restriction of $c_{j0}=-\infty$ and $c_{jK}=+\infty$. Second, in logistic models, c_{jk} and d_j are scaled by the square root of the variance of the logistic distribution, $\pi^2/3$.

The use of parameters of c_{jk} makes it possible for the HRM-SDT to model several rater effects that the HRM (Patz et al., 2002) cannot. For one thing, the HRM-SDT can model raters' central tendency, i.e., preference to not use end categories, by setting c_{j0} and c_{jK} far to the end of the perceptual space (DeCarlo, 2008a; DeCarlo et al., 2011), whereas the HRM (Patz et al., 2002) cannot model this effect since it only has a severity parameter. For another, the HRM-SDT can model raters' preference for specific category scores by a flexible use of c_{jk}. The HRM (Patz et al., 2002), by contrast, can only model raters' overall severity.

The full HRM-SDT

The full HRM-SDT including both levels 1 and 2 is shown in Figure 2.3. This is one type of structural equation model (Bollen, 1989) where the first level is modeled as a latent class SDT model and the second level as a polytomous IRT model (DeCarlo, 2002, 2005, 2008a, 2008b; DeCarlo et al., 2011).

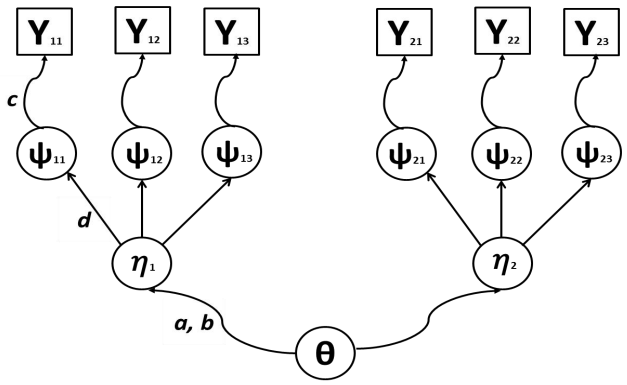


Figure 2.3. Representation of HRM-SDT (DeCarlo et al., 2011)

At the first level, the probability of rater *r*'s scores Y_{jr} is nonlinearly related to her perception Ψ_{jr}, this relationship determined by the response criteria parameter *c*. For the first time, DeCarlo (2008a; 2008b) used curved arrows to indicate a non-linear relation between perception and rater scores. In turn, rater *r*'s perception Ψ_{jr} is linearly related with the latent category η_j of item *j*, this relationship determined by the precision parameter *d*.

At the second level, items' latent categories η_j are used to estimate the examinee proficiency θ via a polytomous IRT model.

2.4 HRM with Covariates (HRMC)

To examine possible influence of factors on rater effects such as bias and variability, researchers (Mariano & Junker, 2007) have incorporated covariates into the HRM and created the HRMC model. They designed a V by Q + S matrix X (see Table 2.2) where V is the number of rows for pseudoraters and Q + S is the number of columns for covariates. Pseudoraters are defined as unique combinations of raters and covariates. The Q columns are indicators of 1 or 0 for each rater, and the S columns contain values for covariates. In Table 2.2, for example, there are four rows for the pseudoraters ($v = 4$); meanwhile, there are four Q columns to indicate which rater each pseudorater involves as well as two S columns to represent values for the two covariates. If it is not intended to examine effects from individual raters, only the S columns are included in the design matrix.

Table 2.2. Design Matrix X for HRMC

Pseudo-rater *v*	Rater1	Rater2	Rater3	Rater4	X_1	X_2
1	1	0	0	0	X_{11}	X_{12}
2	0	1	0	0	X_{21}	X_{22}
3	0	0	1	0	X_{31}	X_{32}
4	0	0	0	1	X_{41}	X_{42}

To model the design matrix, the function of the HRMC model can be represented as

$$p_{\eta kv} = P[Y_{ijv} = k|\eta_{ij} = \eta] \propto \exp\{-\frac{1}{2\Psi_v^2}[k - (\eta + \phi_v)]^2\} \quad (6)$$

where

$p_{\eta kv}$ = probability of pseudorater *v* giving score *k* given ideal rating η,

Ψ_v = variability for pseudorater *v*,

ϕ_v = bias or severity for pseudorater *v*.

For the updated function, the bias effect for pseudorater *v* can be calculated with a linear model as

$$\phi_v = \mathrm{X}_v \boldsymbol{\gamma} \quad (7)$$

where

X_v = the *v*th row of the design matrix for pseudorater *v*,

$\boldsymbol{\gamma}$ = bias vector for the full rating process.

Specifically, the bias vector $\boldsymbol{\gamma} = (\phi_1, \ldots, \phi_Q, \gamma_1, , \ldots, \gamma_S)^{\mathrm{T}}$ contains two components, with the first component representing rater bias effects and the second covariate bias effects.

Likewise, the variability effect (in log scale) can be calculated with a linear model as

$$log\Psi_v^2 = \mathrm{X}_v(\boldsymbol{log\omega^2}) \quad (8)$$

where

Ψ_v^2 = rating variance,

$\boldsymbol{log\omega^2}$ = log-scale rater and covariate rater effects.

Specifically, the variability vector $\boldsymbol{log\omega^2} = (log\Psi_1^2, \ldots, log\Psi_Q^2, log\omega_1^2, \ldots, log\omega_S^2)^{\mathrm{T}}$ contains two components, with the first component representing rater variability effects and the second covariate variability effects.

The HRMC model can be specified in two ways, either as a fixed rating effects model or as a random rating effects model (Mariano & Junker, 2007).

Fixed rating effects model

For the fixed rating effects model, the function can be represented as

$$p_{\eta kv} = P[Y_{ijv} = k|\eta_{ij} = \eta] \propto \exp\{-\frac{1}{2(\exp\{X_v(\log\omega^2)\})}[k-(\eta + X_v\gamma)]^2\}. \quad (9)$$

Comparing Equation 9 with Equation 6, you may find that the original bias ϕ_v and variance Ψ_v^2 have been superseded by the new bias $X_v\gamma$ and the new variance $\exp\{X_v(\log\omega^2)\}$ respectively.

Random rating effects model

For the random rating effects model, the bias effect parameter ϕ_v and variance parameter Ψ_v^2 in Equation 6 are both considered as random variables. The bias effect parameter ϕ_v follows a normal distribution

$$\phi_v \sim N(X_v\boldsymbol{\gamma}, \sigma_\phi^2). \quad (10)$$

The variability effect parameter Ψ_v^2 follows a log-normal distribution

$$log\Psi_v^2 \sim N(X_v\, log\boldsymbol{\omega^2},\ \sigma_\Psi^2). \quad (11)$$

Different from the fixed predictors of $X_v\gamma$ and $X_v(\log\omega^2)$ in Equation 9, the corresponding random predictors have errors $\varepsilon_{\phi,v}$ and $\varepsilon_{\Psi^2,v}$ respectively.

2.5 Latent Class Signal Detection Theory with Covariates (LC-SDTC) Model

There have also been studies incorporating covariates into the HRM-SDT model (Wang, 2012), producing the latent class signal detection theory with covariates (LC-SDTC) model. Compared with the LC-SDT model, the LC-SDTC model incorporates covariates that may influence the rating process and thus various rater effects.

The LC-SDTC model is illustrated in Figure 2.4. Covariates **X**, a vector of variables, are linearly related to latent perceptual categorical variables Ψ, a relationship that is modeled with multiple regression approaches (Wang, 2012). Researchers can also introduce covariates to affect Y, η or θ.

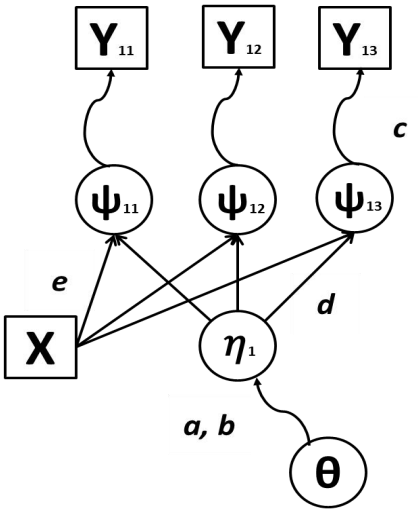


Figure 2.4. Representation of HRM-SDTC (Wang, 2012)

2.6 Extensions of HRM-SDT

The models reviewed so far have studied rater effects from different perspectives and deepened our understanding of how covariates may influence these rater effects. However, there has been little effort to study possible effects of various simplifying assumptions. The present study examines an extension of HRM-SDT where the parameter *d* is not fixed to be equally spaced and examines how this modification of the HRM-SDT may influence the model estimation and fit. This extension makes the latent variable in the new model similar to that in the ordered cluster model or ordinal latent class model (Johnson & Albert, 2006; Uebersax, 1993; van Onna, 2004). But the difference is that the ordered perception model, as the name implies, allows the rater perception rather than the latent essay score variable to be ordinal, whereas the

ordered cluster model or ordinal latent class model are statistical models making the latent score variable ordinal.

The flexible HRM-SDT is represented in Figure 2.5, where *d* is allowed to vary across categories. Compared to the original model in Figure 2.2, the extensions of HRM-SDT have more parameters to be estimated. More generally, with *k* categories, *k*-1 parameters are included in the first level of the hierarchical model.

If in reality the true underlying *d* parameters are unequally spaced, then it seems reasonable to attempt to estimate unequal *d's*. This is equivalent to allowing for different slopes for each item category in Equation 5, similar to the GPCM model in Muraki (1992). But there is no interaction between *d* and *c*, since this will make the model unidentifiable. In Figure 2.5 it is easier for raters to tell category 1 from category 2 than from category 2 from category 3. The present study examines whether or not an extended model can detect this.

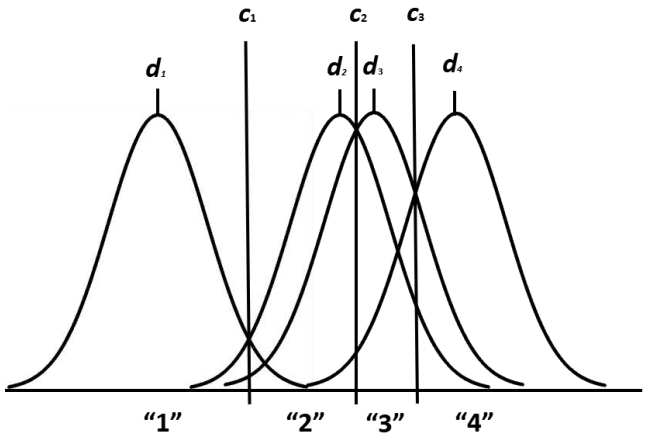


Figure 2.5. Representation of Ordered Perception HRM-SDT

Since the d_1, d_2, d_3, and d_4 notation is used for the ordered perception model, then 'unequal' applies to the differences, $d_2 - d_1$, and $d_3 - d_2$ and so on and not the *d*. In the equal perception model, these differences are all equal to *d*.

Based on the discussion above and the information in Figure 2.5, the latent class SDT unequal perception model can be written as

$$p(Y_j \leq k \mid \eta_m = \eta) = F(c_{jk} - \sum_{m=1}^{M} d_{jm}\eta_m) \quad (12)$$

where

η_m = nominal indicator for latent category m, either 0 or 1,

d_{jm} = precision for rater j and category m of latent perception (DeCarlo & Zhou, 2019).

M is the number of latent categories and response categories so $M=K$. Restrictions are in place that $d_{j1}=0$, $\eta_1=0$, and $d_{j1} \leq d_{j2} \ldots \leq d_{jM}$. Equation 12 has the same two standard assumptions as Equation 5. For a four-class SDT ordered perception rater model, the cumulative probability is $F(c_{jk} - d_{j1}\eta_1 - d_{j2}\eta_2 - d_{j3}\eta_3 - d_{j4}\eta_4)$. The dummy coding for the nominal latent perception variables are as follows.

η_1	η_2	η_3	η_4
0	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Each row shows the dummy coding for that category. Here, the first category (0 0 0 0) is the reference category. In terms of implementation, Croon's (1990) approach was used of imposing ordering through order restrictions on the probabilities (Vermunt & Magidson, 2016). Restrictions on the probabilities are the same in this case as an order restriction on the d's, so Croon's approach and LG were used to implement the parameter restriction.

Equation 12 specifies the cumulative probability model for rater j. The SDT ordered perception model is a restricted latent class model (Clogg, 1995; Dayton, 1998). Assuming J raters and M mutually exclusive latent classes η, the probabilities of the response patterns Y_1—Y_J can be derived by summing over the probability weighted latent classes as

$$
\begin{aligned}
& p(Y_1 = k_1, \ldots, Y_J = k_J) \\
& = \sum_{m=1}^{M} p(Y_1 = k_1, ..., Y_J = k_J, \eta = \eta_m) \\
& = \sum_{m=1}^{M} p(\eta = \eta_m)\, p(Y_1 = k_1, ..., Y_J = k_J \mid \eta = \eta_m) \\
& = \sum_{m=1}^{M} p(\eta = \eta_m) \prod_{j=1}^{J} p(Y_j = k_j \mid \eta = \eta_m), \qquad (13)
\end{aligned}
$$

with $\sum_{m=1}^{M} p(\eta = \eta_m) = 1$ and $\sum_{m=1}^{M} p(Y_j = k_j \mid \eta = \eta_m) = 1$. The last step is based on independence of responses given the latent class.

The SDT is incorporated into the latent class model by obtaining the conditional probability of each response which is the difference of the cumulative probabilities in Equation 12. The conditional probability of each response by each rater is

$$
\begin{aligned}
& p(Y_j = k_j \mid \eta = \eta_m) = F(c_{jk} - \sum_{m=1}^{M} d_{jm}\eta_m) && k_j = 1 \\
& p(Y_j = k_j \mid \eta = \eta_m) = F(c_{jk} - \sum_{m=1}^{M} d_{jm}\eta_m) - F(c_{j(k\text{-}1)} - \sum_{m=1}^{M} d_{jm}\eta_m) && 1 < k_j < K_j \\
& p(Y_j = k_j \mid \eta = \eta_m) = 1 - F(c_{j(k\text{-}1)} - \sum_{m=1}^{M} d_{jm}\eta_m) && k_j = K_j. \qquad (14)
\end{aligned}
$$

So, the full SDT ordered perception rater model is a restricted latent class model having a SDT probability for each rater.

2.7 Diagnostic Measures for Model Fit

Researchers need to cope with the fact that models with more parameters will fit better and can lead to overfitting. The goodness of the fitted model will not generalize to new data. A commonly used measure that suffers from this problem is model deviance which tends to be used to compare performance of nested models. Deviance contains the marginal likelihood of the observed rater scores given model parameter estimates. It can be written as

$$D = -2logP(X|\Phi),$$

where X is the observed rater scores and Φ is the model parameter estimates. LR test is a test of proportional odds. Subtract the D for the fits of the two models and df is the difference in the number of parameters between the two models. Strictly speaking, it is not an LR test because the PME with Bayes constants of 1 was used, but it is close enough to be of interest.

To tackle the overfitting problem of D, statisticians have propose various diagnostic measures such as AIC, BIC, and DIC that not only incorporate D but also use parameter numbers and/or degree of freedom. One of the earliest information criteria proposed to penalize an increase in the number of parameters is AIC.

Akaike information criterion (AIC)

Strictly speaking , AIC (Akaike, 1973) is not a Bayesian measure of model evaluation, but it does penalize for an increased number of parameters. The equation for AIC is

$$AIC = D + 2k,$$

where D is the deviance and k is the number of free parameters to be estimated. Since smaller values of AIC mean better model fit, bigger k increase AIC and thus penalize models with smaller degrees of freedom.

Bayesian information criterion (BIC)

Similar to AIC, BIC (Schwarz, 1978) also penalizes increased parameters in the model, but meanwhile it also penalizes increase number of observations. BIC is a Bayesian approach since it approximates the Bayes factor (Kass & Raftery, 1995). The equation of BIC is

$$BIC = D + k \log(n),$$

where D is the deviance, k is the number of free parameters to be estimated, and n is the number of observations. Since smaller values of BIC mean better model fit, bigger k and/or bigger n increase BIC and thus penalize models with smaller degrees of freedom or those models built on more observations.

Deviance information criterion (DIC)

DIC (Spiegelhalter et al., 2002) is a relatively new information criterion for evaluating model fit. Similar to AIC, DIC only penalizes the degree of freedom. However, different from AIC or BIC that uses exact number of parameters, DIC uses an effective number of parameters that is close to the mode of the likelihood function or the posterior distribution function. The equation of DIC is

$$DIC = D + 2p_D,$$

where p_D is effective number of parameters (Ando, 2011; Gelman & Hill, 2007).

Of course, readers may refer to other means of assessing the fit of a model. To compare non-nested models, researchers may refer to modified versions of AIC or BIC (Burnham & Anderson, 2004; Claeskens & Hjort, 2008). To use methods related to the Bayes factor or marginal likelihood based on MCMC (Chib, 1995; Chib & Jeliazkov, 2001; Gelman & Meng, 1998; Neal, 2001) that have been applied to HRM models, researchers may refer to Mariano (2002). To assess the fit of a model itself, researchers can use cross-validation methods or

posterior predictive check methods (Gelman et al., 2014; Levy, Mislevy, & Sinharay, 2009; Sinharay, Johnson, & Stern, 2006; Zhu & Stone, 2011).

2.8 Estimation Methods

HRM can be estimated via either marginal maximum likelihood (MML; Hombo & Donoghue, 2001), or Bayesian estimation (MCMC; Johnson, 2012; Junker, Patz, & Vanhoudnos, 2012; Patz et al., 2002) methods. For the HRM-SDT model, a partial Bayesian approach posterior mode estimation (PME; DeCarlo et al., 2011) methods can be implemented through the expectation-maximization (EM; Dempster, Laird, & Rubin, 1977; Wu, 1983) algorithm in Latent Gold (Vermunt & Magidson, 2016). The PME is a maximum a posteriori (MAP) estimator which picks the estimate with the highest probability. In the current study, all models were fitted with the PME method.

One advantage of PME is that it can tackle boundary problems. Boundary problems occur when the maximum likelihood estimates (MLE) of some parameters are close to the boundary, "such as obtaining an estimate of a latent class size of zero or unity or obtaining a large or indeterminate estimate of detection (with a large or indeterminate standard error)" (DeCarlo et al., 2011, p. 343). PME solves boundary problems through adding a prior, which penalizes boundary solutions, to the log posterior function and maximizing this function (DeCarlo et al., 2011), so in this sense it is partly Bayesian. The usefulness of PME in dealing with boundary problems has been studied by many researchers (Galindo-Garre & Vermunt, 2006; Gelman et al., 2014; Maris, 1999; Schafer, 1997; Vermunt & Magidson, 2016).

Chapter 3

Methods

The current study focuses on parameter estimation, using bias, absolute percent bias, and mean squared error (MSE) to assess estimation. The percent bias is defined as

$$\%\text{Bi}as = \frac{\hat{\theta} - \theta}{\theta} \times 100,$$

where θ is the simulated true parameter and $\hat{\theta}$ the mean of the 100 parameter estimates. Generally speaking, %bias < 5% is trivial, 5%<%bias<10% moderate, and %bias>10% large (Flora & Curran, 2004). In this study, the absolute percent biases are shown. Note that parameter recovery of large percent biases may not be an issue practically if the bias itself is not large. Or for some cases, even large biases are not problematic. For example, that a true *d* of 6 is estimated as 9 is not practically problematic since *d*'s over 5 all mean great precision in perception.

3.1 Simulation Studies

The simulation design was based on the following research questions. Four groups of simulations, 16 in total, were carried out.

Simulation 1: Generate equal perception model, fit equal perception model

The data were generated from equal perception models and equal perception models were estimated. This type of simulation was conducted in previous studies and results were similar (DeCarlo, 2005, 2008b). Table 3.1 shows the parameters for equal perception models without rater effects where the criteria ware located at the intersection points of adjacent latent perception variables. The *d*'s were between 1 and 5.5 and the *c*'s could be calculated from the formulas in the note under the table.

Table 3.1. Parameters for Equal Perception Model Without Rater Effects

	Raters (*j*)									
	1	2	3	4	5	6	7	8	9	10
d_{j1}	1	2	3	4	5	5.5	4.5	3.5	2.5	1.5
d_{j2}	2	4	6	8	10	11	9	7	5	3
d_{j3}	3	6	9	12	15	16.5	13.5	10.5	7.5	4.5

	Raters (*j*)									
	1	2	3	4	5	6	7	8	9	10
c_{j1}	0.5	1	1.5	2	2.5	2.75	2.25	1.75	1.25	0.75
c_{j2}	1.5	3	4.5	6	7.5	8.25	6.75	5.25	3.75	2.25
c_{j3}	2.5	5	7.5	10	12.5	13.75	11.25	8.75	6.25	3.75

Note
$c_{j1}=1/2d_{j1}$; $c_{j2}= 3/2d_{j1}$; $c_{j3}=5/2d_{j1}$.

Latent-class sizes			
1	2	3	4
0.15	0.35	0.35	0.15

Table 3.2. Parameters for Equal Perception Model with Rater Effects

	Raters (*j*)									
	1	2	3	4	5	6	7	8	9	10
d_{j1}	1	2	3	4	5	5.5	4.5	3.5	2.5	1.5
d_{j2}	2	4	6	8	10	11	9	7	5	3
d_{j3}	3	6	9	12	15	16.5	13.5	10.5	7.5	4.5

	Raters (*j*)									
	1	2	3	4	5	6	7	8	9	10
c_{j1}	0.5	0	2.5	3	4.5	0.75	2.25	0.75	-0.75	0.75
c_{j2}	1.5	2	5.5	6	9.5	6.25	6.75	5.25	3.75	2.25
c_{j3}	2.5	4	8.5	9	14.5	11.75	11.25	9.75	8.25	3.75

Note
Strictness: #3 (+1), #5 (+2) ; Leniency: #2 (-1), #6 (-2) ; Centrality: #8 (-1,0,+1), #9 (-2,0,+2); Extremity: #4 (+1,0,-1).

Latent-class sizes			
1	2	3	4
0.15	0.35	0.35	0.15

Table 3.2 shows the parameters for equal perception models with rater effects where the criteria were shifted to the right or left producing rater effects such as strictness, leniency, extremity, and centrality. Rater effects were reflected by adding or subtracting numbers from *c*'s. The specific rater effects are illustrated in the note under the table. Results of this simulation study provide a baseline for comparison with results from other simulation studies and answer the third research question of how model fit compares for the ordered and equal perception models.

Figure 3.1 shows the rater effects for the equal perception model. Panel A shows the situation where the criteria are shifted to the right of the optimal point, this rater has lower probabilities of assigning higher scores, so this rater is strict. Panel B shows the situation where the criteria are shifted to the left and so this rater is lenient. Panel C shows the situation where the lower criterion is shifted to the left whereas the higher criterion is shifted to the right. This rate has higher probabilities of assigning middle scores, so this rater has centrality effect. Panel D shows the situation where the lower criterion is shifted to the right whereas the higher criterion is shifted to the left. This rate has lower probabilities of assigning middle scores, so this rater has extremity effect.

Simulation 2: Generate ordered perception model, fit ordered perception model

Table 3.3 shows the parameters for ordered perception models without rater effects.

Figure 3.2 shows examples of unequal distances in the ordered perception model where adjacent distributions at different locations are close together. Panel A shows the situation where this rater cannot distinguish latent categories 1 and 2. Panel B shows the situation where this rater cannot distinguish latent categories 2 and 3. Panel C shows the situation where this rater

Table 3.3. Parameters for Ordered Perception Model Without Rater Effects

	Raters (*j*)									
	1	2	3	4	5	6	7	8	9	10
d_{j1}	1	1	3	1	5	5.5	1	1	2.5	1.5
d_{j2}	2	3	6	5	6	11	5.5	4.5	3.5	3
d_{j3}	3	5	7	6	11	16.5	6.5	8	6	4

Note
L: #2 (-1,0,0), #8 (-2.5,0,0); M: #5 (0,-4,0), #9 (0,-1.5,0); H: #3 (0,0,-2), #10 (0,0,-0.5);
E: #4 (-3,0,-3), #7 (-3.5,0,-3.5). L, M, H, and E mean low, middle, high, and end locations.

	Raters (*j*)									
	1	2	3	4	5	6	7	8	9	10
c_{j1}	0.5	0.5	1.5	0.5	2.5	2.75	0.5	0.5	1.25	0.75
c_{j2}	1.5	2	4.5	3	5.5	8.25	3.25	2.75	3	2.25
c_{j3}	2.5	4	6.5	5.5	8.5	13.75	6	6.25	4.75	3.5

Note
c_{j1}=1/2d_{j1}; c_{j2}=1/2 (d_{j1}+d_{j2});c_{j3}=1/2(d_{j2}+d_{j3}).

Latent-class sizes			
1	2	3	4
0.15	0.35	0.35	0.15

Table 3.4. Parameters for Ordered Perception Model with Rater Effects

	Raters (*j*)									
	1	2	3	4	5	6	7	8	9	10
d_{j1}	1	1	3	1	5	5.5	1	1	2.5	1.5
d_{j2}	2	3	6	5	6	11	5.5	4.5	3.5	3
d_{j3}	3	5	7	6	11	16.5	6.5	8	6	4

Note
L: #2 (-1,0,0), #8 (-2.5,0,0); M: #5 (0,-4,0), #9 (0,-1.5,0); H: #3 (0,0,-2), #10 (0,0,-0.5);
E: #4 (-3,0,-3), #7 (-3.5,0,-3.5).

	Raters (*j*)									
	1	2	3	4	5	6	7	8	9	10
c_{j1}	0.5	-0.5	2.5	1.5	4.5	0.75	0.5	-0.5	-0.75	0.75
c_{j2}	1.5	1	5.5	3	7.5	6.25	3.25	2.75	3	2.25
c_{j3}	2.5	3	7.5	4.5	10.5	11.75	6	7.25	6.75	3.5

Note
Strictness: #3 (+1), #5 (+2) ; Leniency: #2 (-1), #6 (-2) ; Centrality: #8 (-1,0,+1), #9 (-2,0,+2);
Extremity: #4 (+1,0,-1).

Latent-class sizes			
1	2	3	4
0.15	0.35	0.35	0.15

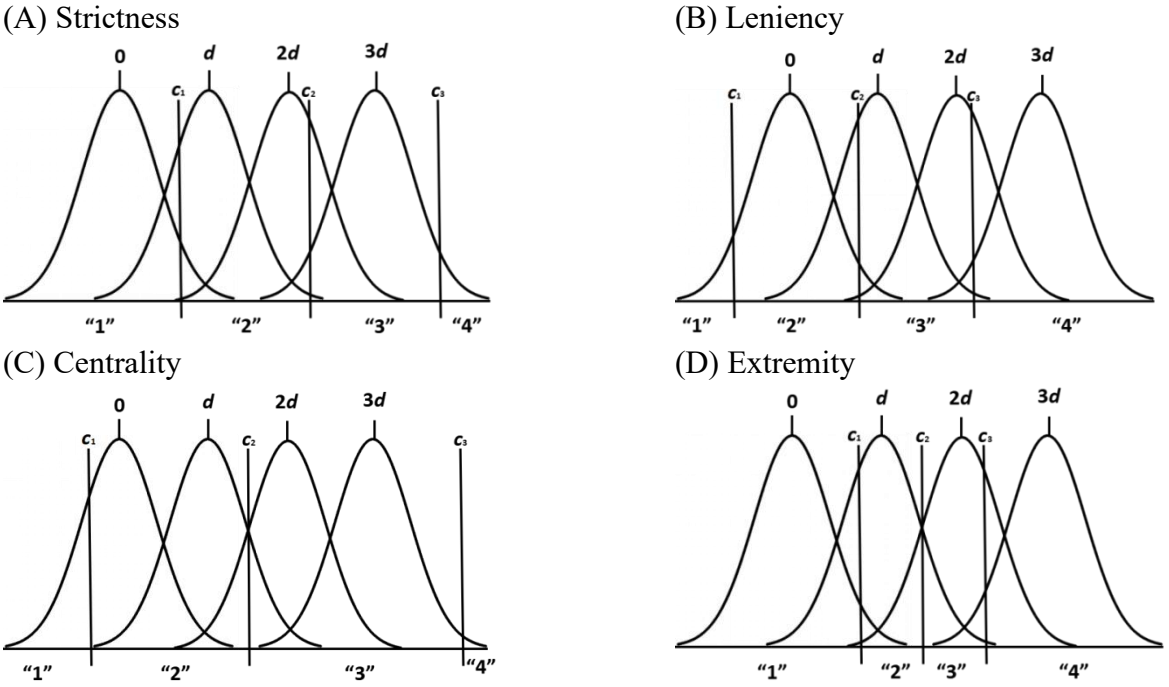


Figure 3.1. Representation of Rater Effects in Equal Perception Models.

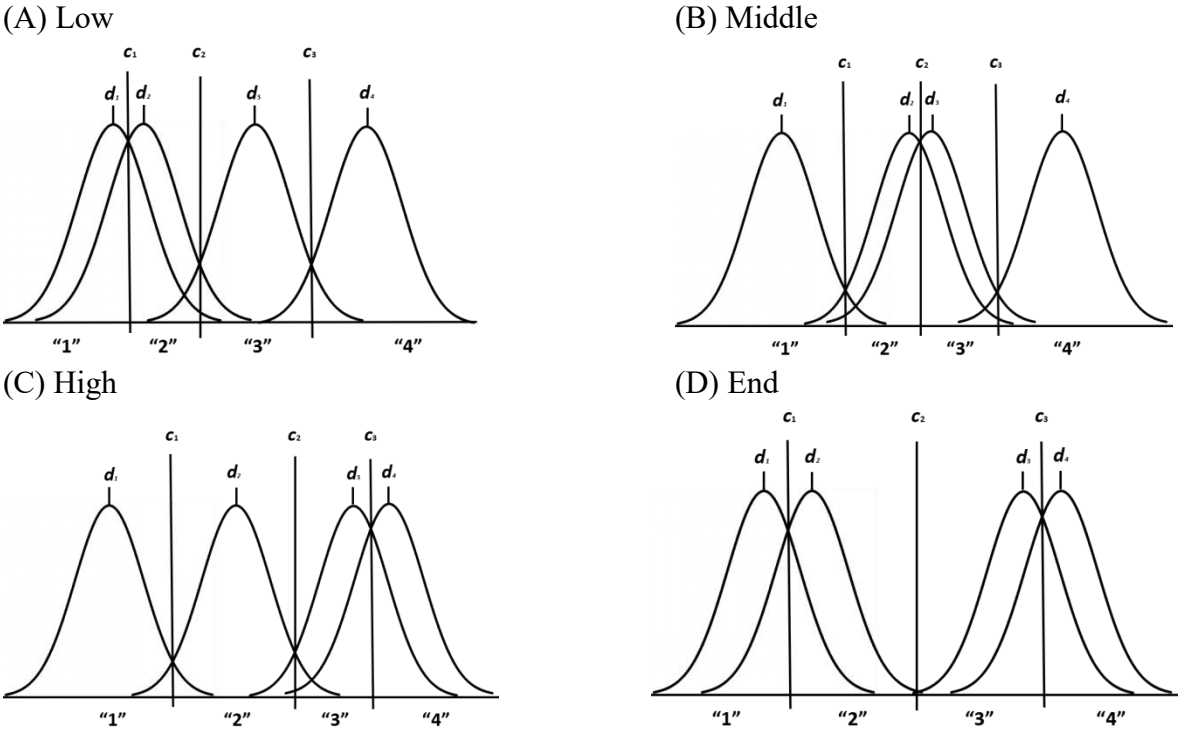


Figure 3.2. Representation of Unequal Distances in Ordered Perception Models Without Rater Effects. One panel respectively represents one place where adjacent distributions are difficult to distinguish.

cannot distinguish latent categories 3 and 4. Panel D shows the situation where this rater can distinguish neither between latent categories 1 and 2 nor between 3 and 4 but can tell 2 from 3.

Table 3.4 shows the parameters for ordered perception model with rater effects. The true adjacent perception distributions had unequal *d*'s between 1 and 5.5. Rater effects were added as in simulation 1. Ordered perception models were fitted. Results answer the first research question of to what extent model parameters can be recovered for the ordered perception model.

Simulation 3: Generate ordered perception model, fit equal perception model

The data in simulation 2 were fitted with equal perception models. Results answer the second and third research questions of to what extent fitting wrong models will affect parameter recovery and how model fit compares for the ordered and equal perception models.

Simulation 4: Generate equal perception model, fit ordered perception model

The data in simulation 1 were fitted with ordered perception models. Results answer the second and third research questions of to what extent fitting wrong models will affect parameter recovery and how model fit compares for the ordered and equal perception models.

Complete versus incomplete data

Simulations in this study used a fully-crossed design, where each essay is scored by each rater, and balanced incomplete block (BIB) design, where each essay is scored by the same number of raters and each rater rates the same number of essays. Since for 10 raters a strict BIB design should involve 1,080 students, the current study which had 1,000 students is an approximate BIB design. As shown in Table 3.5 (DeCarlo, 2010), there are 45 rater pairs, each pair scores 24 essays, each rater scores 216 essays, and in total 1,080 essays are scored. Every possible combination of raters is used. For example, the rater pair of 1 and 2 or the rater pair of 2 and 3 scores the same 24 essays. For BIB design, the same 100 corresponding data sets without

Table 3.5. Balanced Incomplete Block (BIB) Design, 45 Rater Pairs, 24 per Pair

	Rater										
	1	2	3	4	5	6	7	8	9	10	Total
	24	24									
	24		24								
	24			24							
	24				24						
	24					24					
	24						24				
	24							24			
	24								24		
	24									24	
		24	24								
		24		24							
		24			24						
		24				24					
		24					24				
		24						24			
		24							24		
		24								24	
			24	24							
			24		24						
			24			24					
			24				24				
			24					24			
			24						24		
			24							24	
				24	24						
				24		24					
				24			24				
				24				24			
				24					24		
				24						24	
					24	24					
					24		24				
					24			24			
					24				24		
					24					24	
						24	24				
						24		24			
						24			24		
						24				24	
							24	24			
							24		24		
							24			24	
								24	24		
								24		24	
									24	24	
											1080
Total/Rater	216	216	216	216	216	216	216	216	216	216	

and with rater effects in the fully-crossed design were used, only that scores were randomly deleted to create the BIB design with 80% of scores missing. Each essay was rated by two raters. Seven raters rated 199 essays, two raters 200 essays, and one rater 207 essays.

Relative criterion plot

Figure 3.1 is a relative criterion plot for the equal perception model showing the various rater effects generated in the four simulation studies. For the equal perception model, the relative response criterion is calculated as each estimated criterion parameter divided by the estimated distance between the highest and lowest latent perception distributions for each rater, $c_{jk}/[(K-1)\times d_j]$ (DeCarlo, 2005). Then, it is possible to compare rater effects among each rater. Raters 1, 7, and 10 have no rater effects since their estimated criteria are located on the lines. Raters 8 and 9 have centrality effects since their first criteria are below the optimal line and their third criteria above the line. Since the criteria are thresholds of assigning scores, these two raters will never give scores of 1 or 4. Rater 4 has extremity effect since her first criteria are above the optimal line and her third criteria below the line. Raters 2 and 6 are lenient since their three criteria are all below the lines. Raters 3 and 5 are strict since their three criteria are all above the lines.

Data generation

The data were simulated with modified SAS macros by DeCarlo (2005). 10 raters discriminated between four latent classes and assigned a score of 1-4. The latent class sizes followed an approximately normal distribution, in consistency with the results from analyses of real-world data (DeCarlo, 2008b). The logistic-model values of *d* from 0.5–5.5 were used, which means moderate to excellent rater precision (DeCarlo, 2008b). Distance less than 0.5 is small, over 2 big, and around 1 medium (DeCarlo, 2002, 2005). Without rater effects, the criteria are located at the interaction of adjacent latent perception distributions. Each condition had a sample

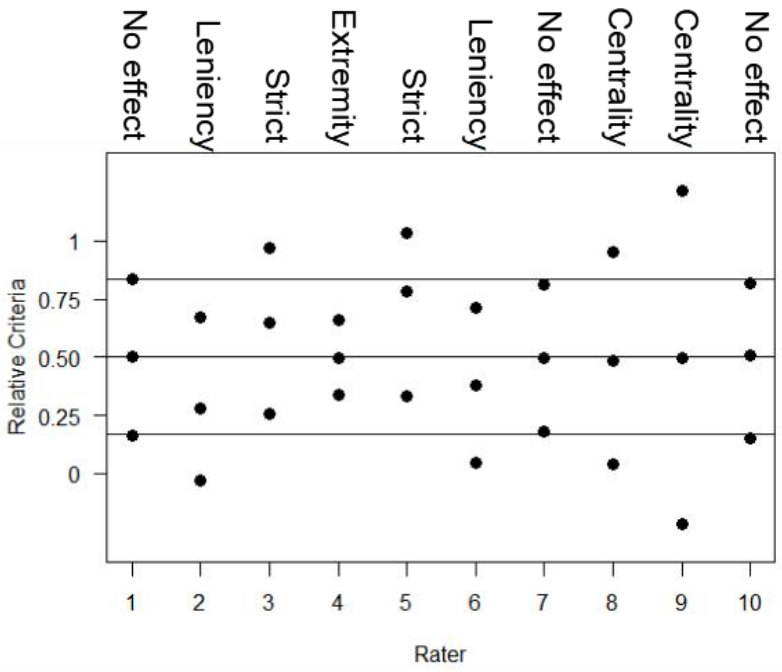


Figure 3.3. Relative Criteria Parameters for a 4-class Equal Perception Model. The plot shows relative response criteria for each rater as filled circles and optimal location points as lines. Names of rater effects are shown on top.

size of 1,000 and 100 datasets. The process of data generation followed three steps as in DeCarlo (2008b):

1. Generate the values for the latent variable η (0, 1, 2 and 3) with a multinomial distribution. The probability for each latent category followed the size of each latent class.
2. Generate cumulative response probabilities for raters and response categories with a logistic distribution for *F*. Plug η obtained in step 1 together with population parameters *crkj* and *drj* into Equation 5 or *crkj* and *drkj* into Equation 12.
3. Generate an observed response. The probabilities generated in step 2 were compared to the value of a uniform random variable sampled from 0–1. The response was assigned 1, if the value of the uniform random variable was smaller than or equal to the probability of the lowest response category; was assigned 2, if the value was

greater than the probability of the lowest response category but smaller than or equal to the probability of the second response category; and so on.

Model estimation

Latent Gold 5.1 (Vermunt & Magidson, 2016) was used to fit the equal and ordered perception models. First, simulation studies showed that Version 4.5 had good performance for latent class SDT models (DeCarlo, 2008b). Second, Version 5.1 allows for numerous models and constraints. Third, for latent discrete variables, the algorithm in Latent Gold converges hundreds of times faster than Mplus. Latent Gold starts with the EM algorithm and shifts to the Newton-Raphson algorithm when estimate is in the neighborhood of the ML value (Vermunt & Magidson, 2016). Fourth, Bayes methods can handle missing data and boundary parameters such as large *d*'s. All models were estimated with the PME method.

Note that label switching (McLachlan & Peel, 2000) is a problem that should be tackled before summarizing the results from estimation. It is arbitrary to assign latent classes as 0, 1, 2, 3 or 3, 2, 1, 0. Since the order of the estimated latent classes is arbitrary, it is possible to have two solutions of *d* that have the same log likelihood values but reversed signs. To address this problem, a (+) was added before each latent class in the equations of Latent Gold. The plus imposes a monotonicity constraint on the probability of each latent category so that $d_1 \leq d_2 \leq d_3$.

3.2 Empirical Study

Essay scores of 2,350 test taker from a large-scale language test were analyzed to compare the ordered perception model to the equal perception model. The essays were scored by 27 raters using scores of 1-5, each was rated by two raters, and each rater rated from 61 to 354 essays.

Chapter 4

Results

This chapter shows results for the four simulation studies and the real data analysis. The first five sections discuss results for simulated data both without and with rater effects, both for the fully-crossed design and the BIB design. Also shown is how information criteria are useful in picking the correct models. Section 4.6 discusses results for analyses of real data.

4.1 Simulation 1: Equal Perception Data, Fit Equal Perception Model

Without rater effects

Table A1.1 shows parameter recovery of *d's* and *c's* over 100 replications of the fully-crossed design without rater effects. The recovery was excellent, with all the percent biases below 3% and all the MSE's below 0.3. The recovery of the latent class sizes was also excellent, with all the percent biases below 1% and all the MSE's <0.001.

For the equal perception model, the relative response criterion is calculated as each estimated criterion parameter divided by the estimated distance between the highest and lowest latent perception distributions for each rater, $c_{jk}/[(K\text{-}1)\times d_j]$ (DeCarlo, 2005). Then, it is possible to compare rater effects among each rater.

Figure 4.1 shows the distance and relative criteria parameters for the 4-class model. Panel (A) shows a plot of the estimated discrimination parameters with an error bar for each rater which is almost too small to see. The error bar is calculated as $\pm 2 \times \frac{SD}{\sqrt{100}}$. It is easy to compare the relative discrimination abilities among each rater and to see that the SE's of all raters were

negligible. Panel (B) shows relative response criteria for each rater where there were no rater effects as the SE's of all points overlapped with the lines or optimal locations.

(A) (B)

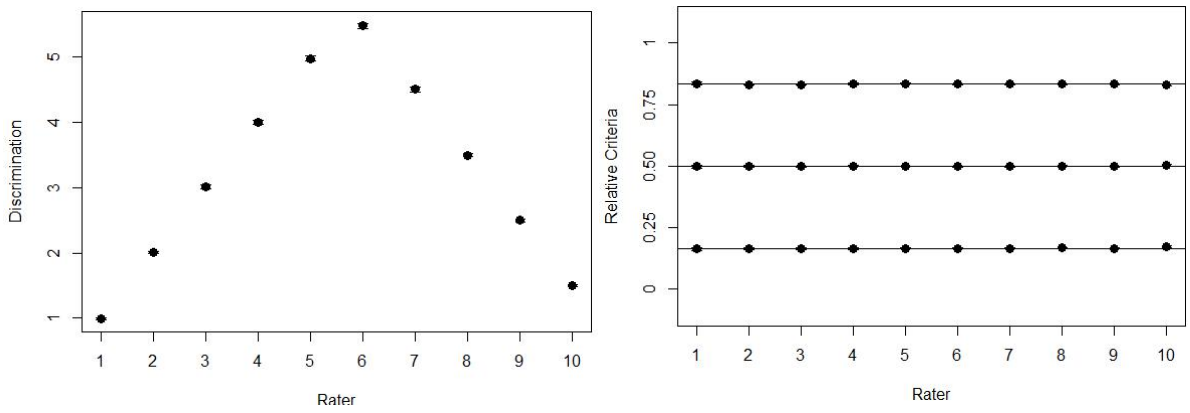


Figure 4.1. Fully Crossed Design, Distance and Criteria Parameters for a 4-class Equal Perception Model Without Rater Effects, Fit Equal Perception Model. Panel (A) shows discrimination parameters for each rater and SE bars. Panel (B) shows relative response criteria for each rater and SE bars and optimal location points as lines.

Table B1.1 shows the parameter recovery of *d's* and *c's* over 100 replications of the BIB design without rater effects. Compared with its fully-crossed counterpart, the quality of parameter recovery was a little worse. Although the first criteria of *c*'s, e.g., c_{11}, c_{31}, c_{41}, tended to be underestimated by 10%-30%, the biases of other parameters were mostly trivial and all MSE's were below 5. For estimates of latent class sizes, the first and fourth classes were overestimated by about 15%, whereas the middle classes were underestimated by slightly over 5%. All MSE's were around 0.001. Therefore, missing data to some extent degraded parameter estimation.

Figure 4.2 shows the distance and relative criteria parameters. Compared with Figure 4.1, the SE's were slightly larger, and most SE's were negligible. Panel (B) shows that there were no rater effects as the SE's of all points overlapped with or were close to the lines or optimal

locations. Therefore, although missing data increased the percent biases of *d*'s and *c*'s and the SE's of *d*'s, they had little effects on detection of rater effects.

(A) (B)

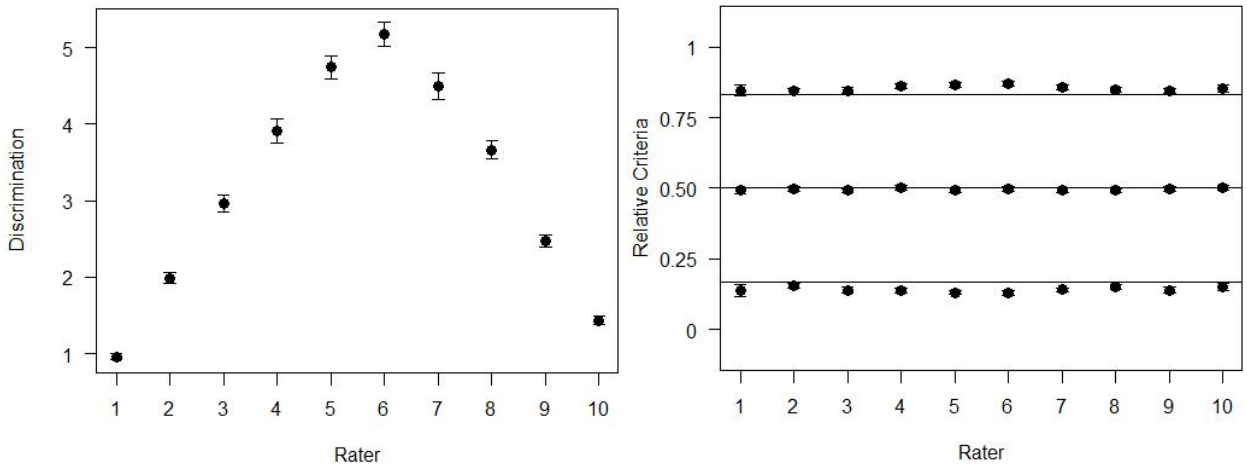


Figure 4.2. BIB Design, Distance and Criteria Parameters for a 4-class Equal Perception Model Without Rater Effects, Fit Equal Perception Models. Panel (A) shows discrimination parameters for each rater and SE bars. Panel (B) shows relative response criteria for each rater, SE bars, and optimal location points as lines.

With rater effects

Table A1.2 shows the parameter recovery of *d's* and *c's* over 100 replications of the fully-crossed design with rater effects. The recovery of parameters was excellent, similar to that of the fully-crossed simulation without rater effects, with all the percent biases below 3%. Note that a dash was used to supersede the infinity percent bias of the estimate of c_{21} which occurred because of a 0 true value. Dashes were also used for other percent biases of true values of 0. The MSE's were mostly below 0.5, except for the largest c_{53} which was slightly over 1. The recovery of the latent class sizes was also excellent, as the percent biases were all below 2% and the MSE's were all <0.001. Similar parameter recovery between without and with rater effects means that adding rater effects in the fully-crossed design had little effect on estimation of model parameters and latent class sizes.

Figure 4.3 shows the distance and relative criteria parameters for the 4-class model. Panel (A) shows that compared with the fully-crossed simulation without rater effects in Figure 4.1, the SE's were a little larger. Panel (B) shows that all the generated rater effects were perfectly caught. For example, either Rater 8's or Rater 9's obtained first relative criterion was lower than the optimal first relative criterion and both of their obtained third relative criteria were high than the optimal third relative criteria, meaning that those two raters had centrality effects. They were more likely to assign scores 2 and 3.

(A) (B)

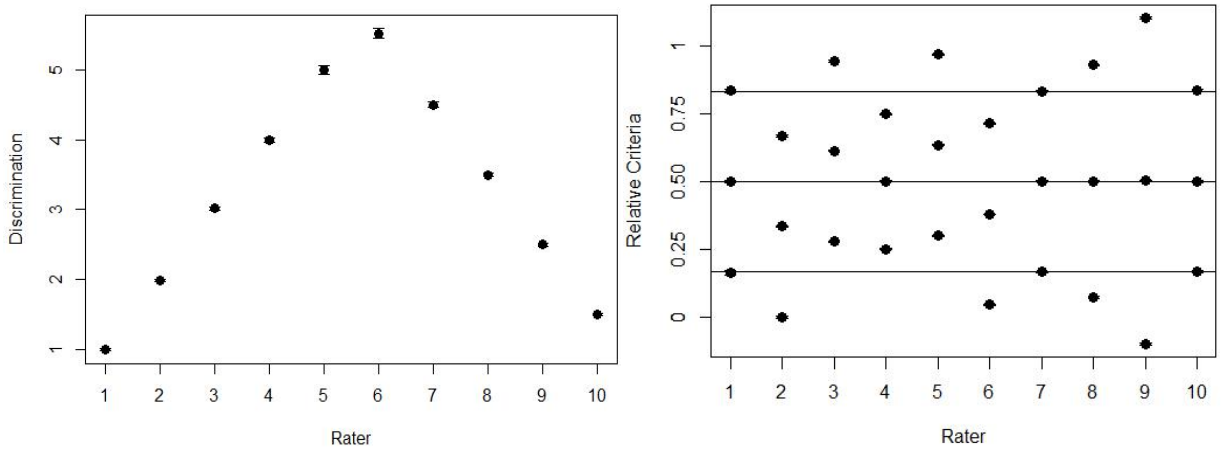


Figure 4.3. Fully Crossed Design, Distance and Criteria Parameters for a 4-class Equal Perception Model with Rater Effects, Fit Equal Perception Model. Panel (A) shows discrimination parameters for each rater and SE bars. Panel (B) shows relative response criteria for each rater and SE bars and optimal location points as lines.

Table B1.2 illustrates the parameter recovery of *d's* and *c's* over 100 replications of the BIB design with rater effects. Compared with its fully-crossed design counterpart, the quality of parameter recovery was a little worse. Although the first criteria of *c*'s, e.g., c_{11}, c_{41}, c_{51}, tended to be underestimated by 10%-60% and boundary *d*'s tended to be underestimated by around 20%, the biases of other parameters were mostly trivial. The MSE's were mostly below 1, except for the largest parameters. For estimates of latent class sizes, the first and fourth classes were

overestimated by about 10%-15%, whereas the middle classes were underestimated by about 5%. The MSE's were around 0.001. Once again, this comparison shows that missing data somewhat biased the estimation of parameters.

Figure 4.4 shows the distance and relative criteria parameters for the 4-class model of the BIB design with rater effects. Panel (A) shows that compared with Figure 4.2, the SE's were slightly larger, which means that for BIB design adding rater effects had little effect on SE's of parameter estimation. Raters 1 and 10 had the smallest standard errors whereas Raters 3-8 had the largest. Raters 5 and 6 had large, negative biases for their *d*'s, and these two raters had the highest true *d*'s and relatively large criteria shifts—up 2 for d_5 and down 2 for d_6. So, large rater effects can bias the *d*'s a little. Panel (B) shows that all rater effects were determined. Once again, it was found that missing data had little effect on detection of rater effects.

(A) (B)

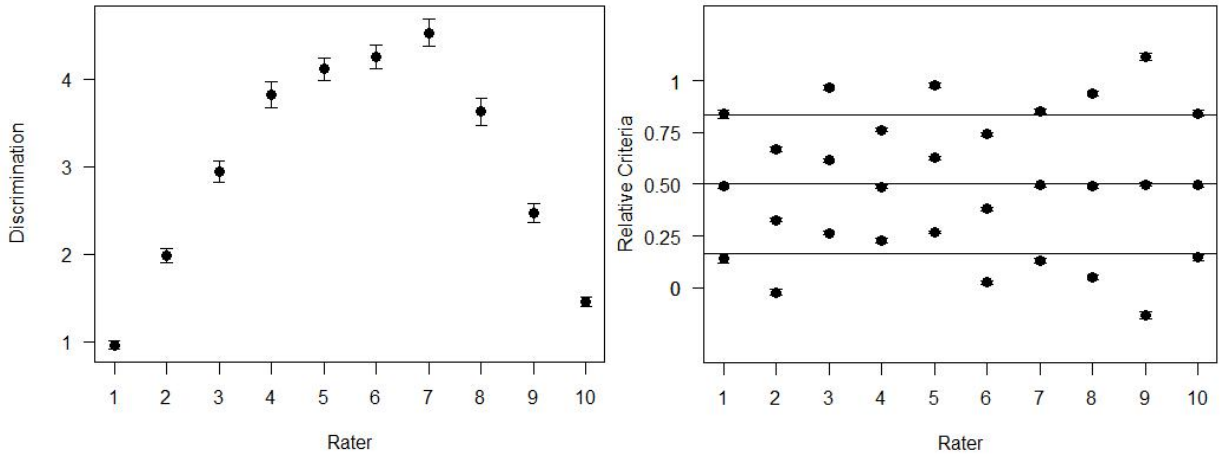


Figure 4.4. BIB Design, Distance and Criteria Parameters for a 4-class Equal Perception Model with Rater Effects, Fit Equal Perception Models. Panel (A) shows discrimination parameters for each rater and SE bars. Panel (B) shows relative response criteria for each rater, SE bars, and optimal location points as lines.

Summary

This section shows that it was easy to recover equal-perception model parameters for equal-perception data, with fully-crossed design generating better estimates than BIB design. Data without rater effects tended to have slightly better parameter recovery and SE than data with rater effects. Plots were useful to detect rater effects despite simulation design.

4.2 Simulation 2: Ordered Perception Data, Fit Ordered Perception Model

Without rater effects

Table A2.1 shows the parameter recovery of *d's* and *c's* over 100 replications of the fully-crossed design without rater effects. The recovery was excellent, with all but one of the percent biases below 5% and the MSE's almost all below 0.5. The recovery of the latent class sizes was also excellent, with percent biases all below 2% and the MSE's all <0.001.

For ordered perception models, the relative response criteria were calculated as each estimated criterion parameter *c* divided by the largest estimated *d* for each rater, $c_{jk}/d_{j(K-1)}$, thus making it possible to detect and compare the rater effects among each rater. Same as the equal perception model, the denominator is the distance between the lowest and highest perception distributions. However, the optimal locations are not on the same line. Each rater has their own set of locations for their optimal locations.Since the relative criteria in the ordered perception model do not lie on the same line, open triangles were used to play the same role as that of the lines in the equal perception plot. The positions of the open triangles were calculated from the estimated *d*'s, located at the optimal intersection points of adjacent latent perception distributions, whereas the filled circles were obtained relative criteria from fitting unequal perception models. By comparing the positions of the filled circles with those of the open triangles, it is easy to

detect various rater effects. For example, if for one rater a circle is above the corresponding triangle, then this rater tends to be strict on assigning this score.

Figure 4.5 shows the relative criteria parameters for each rater. Obviously, there were no rater effects as all circles overlapped with triangles.

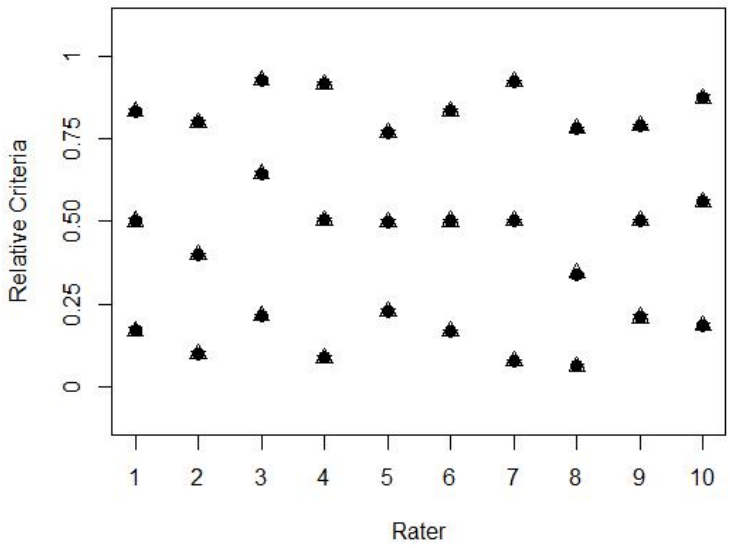


Figure 4.5. Fully Crossed Design, Criteria Parameters for a 4-class Ordered Perception Model Without Rater Effects, Fit Ordered Perception Model. The plot shows relative response criteria for each rater and SE bars, with open triangles as predicted relative criteria from *d* estimates and filled circles as obtained relative criteria estimates.

Table B2.1 shows the parameter recovery of *d's* and *c's* over 100 replications of the BIB design without rater effects. Compared with its fully-crossed design counterpart, the quality of parameter recovery was a little worse. Although some *d*'s and *c*'s tended to be underestimated by 10%-50%, the biases of other parameters were mostly trivial. The MSE's were mostly below 5 except for the largest *d*'s and *c*'s. For estimates of latent class sizes, the first and fourth classes were overestimated by about 20%-30%, whereas the middle classes were underestimated by slightly over 10%. The MSE's were around 0.003. Therefore, missing data to some extent biased the estimation of parameters.

Figure 4.6 shows the relative criteria parameters for each rater. As illustrated, all circles overlapped with or were close to the triangles, except that mild artificial centrality effects were created for Raters 5 and 6.

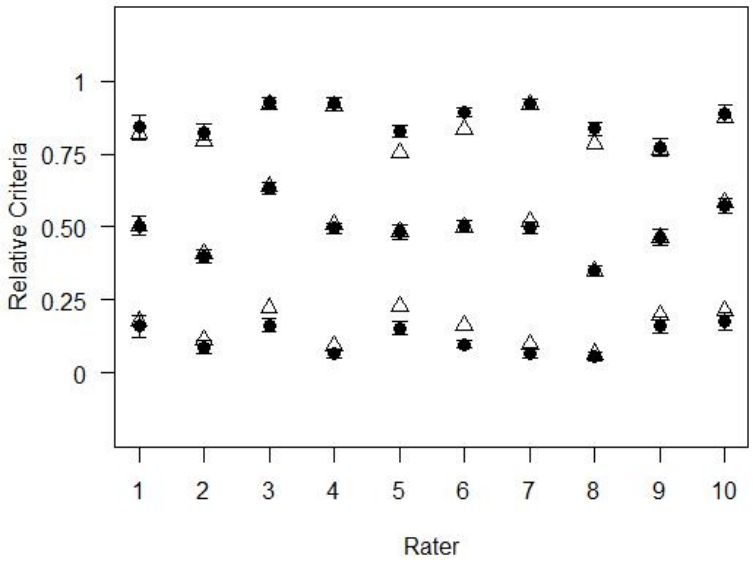


Figure 4.6. BIB Design, Criteria Parameters for a 4-class Ordered Perception Model Without Rater Effects, Fit Ordered Perception Models. The plot shows relative response criteria for each rater and SE bars, with open triangles as predicted relative criteria from *d* estimates and filled circles as obtained relative criteria estimates.

With rater effects

Table A2.2 shows the parameter recovery of *d's* and *c's* over 100 replications of the fully-crossed design with rater effects. Similar to the fully-crossed design without rater effects, the recovery of parameters was excellent, with all but one percent bias below 5% and the MSE's all below 3. The recovery of the latent class sizes was also excellent, as the percent biases were all below 2% and the MSE's were all <0.001. Similar parameter recovery between without and with rater effects mean that adding rater effects in the fully-crossed design did not affect estimation of parameters and latent class sizes.

Figure 4.7 shows the relative criteria parameters for the 4-class model. As shown, all the generated rater effects were determined. For example, Rater 8's or Rater 9's obtained first criteria were lower than their respective optimal first relative criteria, and meanwhile, their obtained third criteria were higher than their respective optimal third relative criteria, meaning that those two raters had centrality effects. They were more likely to assign scores 2 and 3. On the other hand, no artificial rater effects were created for Raters 1, 7, or 10.

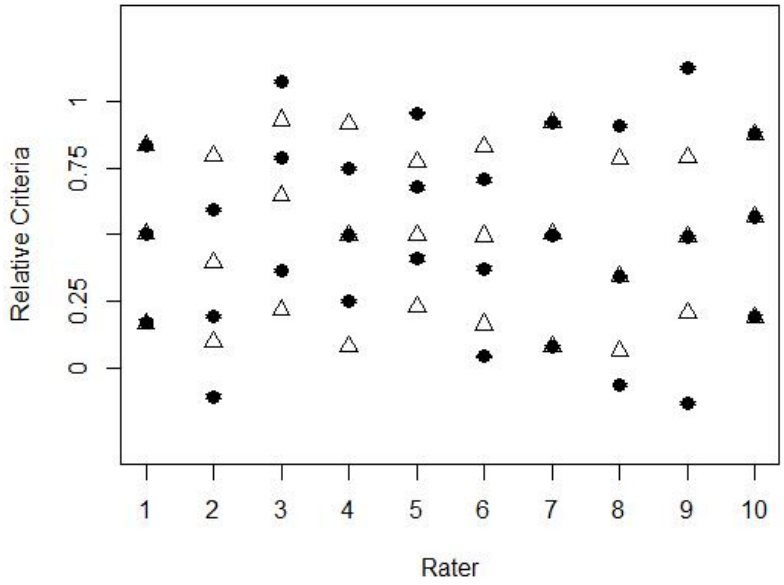


Figure 4.7. Fully Crossed Design, Criteria Parameters for a 4-class Ordered Perception Model with Rater Effects, Fit Ordered Perception Model. The plot shows relative response criteria for each rater and SE bars, with open triangles as predicted relative criteria from *d* estimates and filled circles as obtained relative criteria estimates.

Table B2.2 shows the parameter recovery of *d's* and *c's* over 100 replications of the BIB design with rater effects. Compared with its fully-crossed design counterpart, the quality of parameter recovery appeared worse. Most *d*'s and *c*'s tended to have biases around 30%-40% and the MSE's were mostly below 5 except for the largest *d*'s and *c*'s. But biases of some *d*'s suffered from carry-over effects from biases of d_{r1}. If the parametrization of *d* is changed to the distance between adjacent distributions, then most biases would decrease dramatically. For

estimates of latent class sizes, the first and fourth classes were overestimated by about 30%-50%, whereas the middle classes were underestimated by about 15%. The MSE's were between 0.005 and 0.010. So, missing data biased the estimation of parameters, and the degree of bias was much larger for ordered perception models than for equal perception models.

Figure 4.8 shows the relative criteria parameters for the 4-class model. Nearly all rater effects were determined. A comparison with Figure 4.7 indicates that missing data had trivial effects on detection of rater effects.

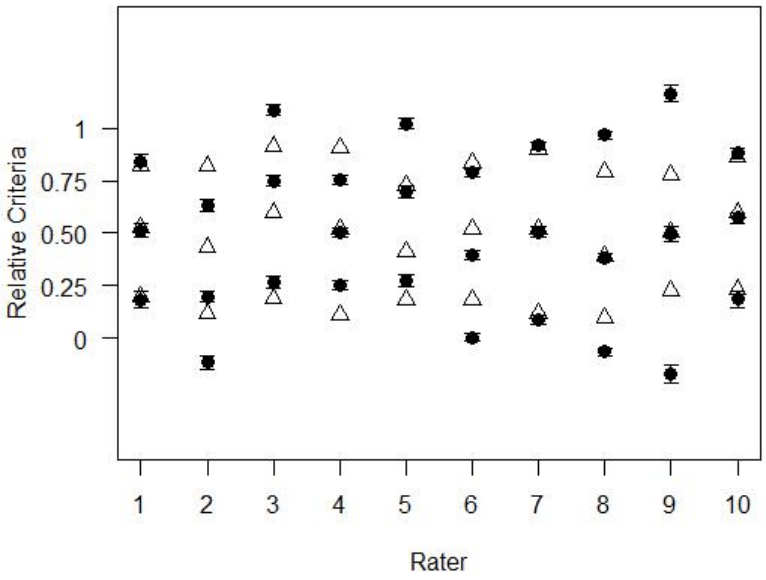


Figure 4.8. BIB Design, Criteria Parameters for a 4-class Ordered Perception Model with Rater Effects, Fit Ordered Perception Models. The plot shows relative response criteria for each rater and SE bars, with open triangles as predicted relative criteria from *d* estimates and filled circles as obtained relative criteria estimates.

Summary

This section shows that, for ordered perception models, parameter recovery for fully-crossed design was much better than for BIB design. Also, data without rater effects had slightly better parameter recovery than data with rater effects. Plots were useful to detect rater effects

despite simulation design. Therefore, the large percent biases for some raters were practically not an issue.

This simulation can answer the first research question: To what extent can model parameters be recovered for the ordered perception model? Results show that fitting correct ordered perception models gave excellent parameter recovery and latent class size estimation. However, the BIB design tended to distort estimates of some parameters in different degrees and created mild artificial centrality effects for some raters.

4.3 Simulation 3: Ordered Perception Data, Fit Equal Perception Model

Without rater effects

Table A3.1 shows the parameter recovery of *d's* and *c's* over 100 replications of the fully-crossed design without rater effects. Fitting the wrong model does not expect recovery of *d*'s. The recovery of the latent class sizes was excellent, with percent deviations all below 2%. Figure 4.9 shows little rater effects as all SE bars were close to the three lines.

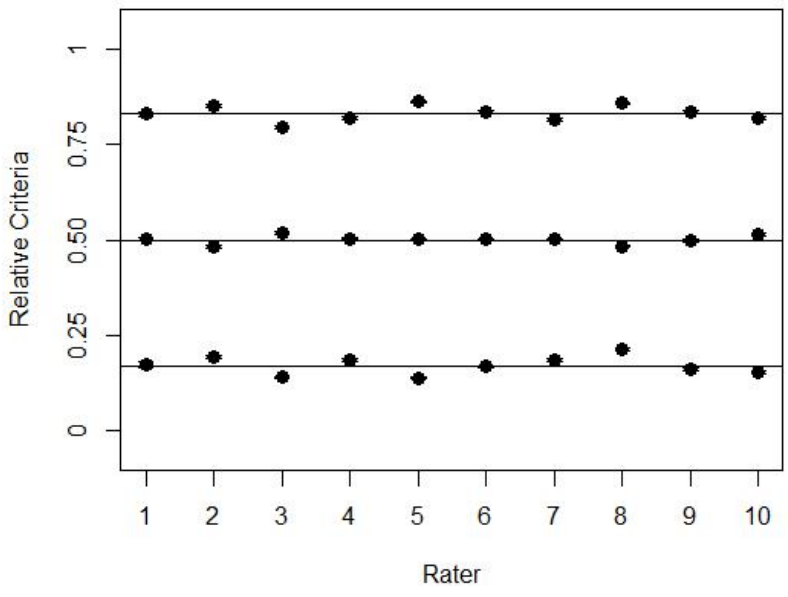


Figure 4.9. Fully Crossed Design, Criteria Parameters for a 4-class Ordered Perception Model Without Rater Effects, Fit Equal Perception Model. The plot shows relative response criteria for each rater and SE bars.

Table B3.1 shows the parameter recovery of *d's* and *c's* over 100 replications of the BIB design without rater effects. Compared with its fully-crossed design counterpart, the quality of parameter recovery was a little worse. Most parameters were deviated by below 50%, with some first *d*'s and *c*'s by 100%-200%. For estimates of latent class sizes, the first and fourth classes were deviated by about 40%-60%, whereas the middle classes by over 20%. Same as before, missing data to some extent degraded the estimation of parameters, especially latent class sizes.

Figure 4.10 shows that detection of rater effects was reasonable as most SE bars overlapped with the three lines. However, some mild centrality effects were created for some raters, such as Rater 5 and Rater 6. This is consistent with large percent deviations of parameter estimates of some boundary *c*'s.

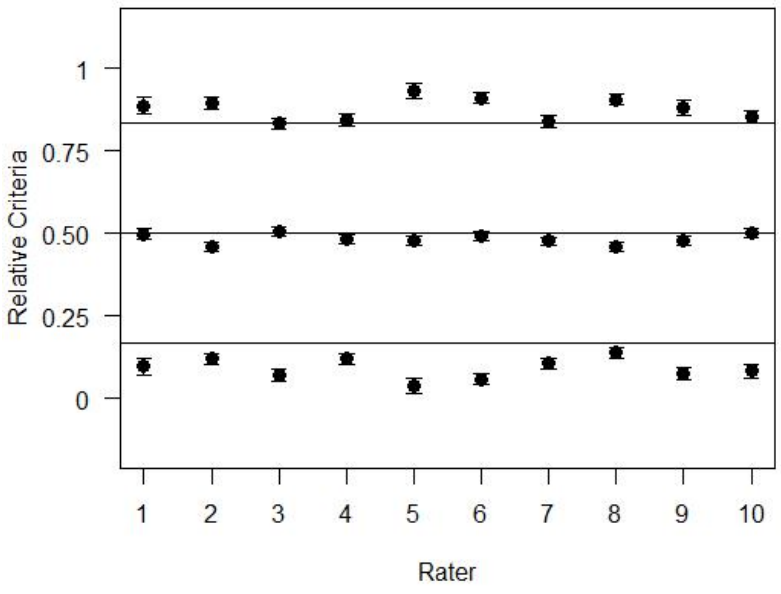


Figure 4.10. BIB Design, Criteria Parameters for a 4-class Ordered Perception Model Without Rater Effects, Fit Equal Perception Models. The plot shows relative response criteria for each rater, SE bars, and optimal location points as lines.

With rater effects

Table A3.2 shows the parameter recovery of *d's* and *c's* over 100 replications of the fully-crossed design with rater effects. Again, fitting the wrong model does not expect recovery of *d*'s. Figure 4.11 shows that all rater effects were determined, which means that fitting wrong models might not affect detection of rater effects.

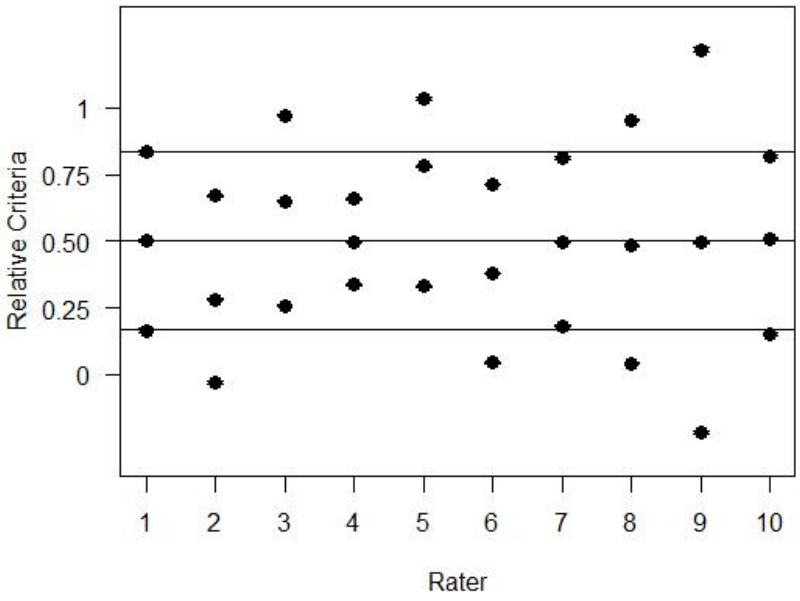


Figure 4.11. Fully Crossed Design, Criteria Parameters for a 4-class Ordered Perception Model with Rater Effects, Fit Equal Perception Models. The plot shows relative response criteria for each rater and SE bars.

Table B3.2 shows the parameter recovery of *d's* and *c's* over 100 replications of the BIB design with rater effects. Compared with its fully-crossed design counterpart, the quality of parameter recovery was a little worse. Most parameters were deviated below 50%, with some first *d*'s and *c*'s by 100%-200%. For estimates of latent class sizes, the first and fourth classes were overestimated by about 70%-100%, whereas the middle classes were underestimated by 30%-40%. Therefore, missing data to some extent degraded the estimation of parameters, especially latent class sizes. Figure 4.12 shows that rater effects were reasonably determined.

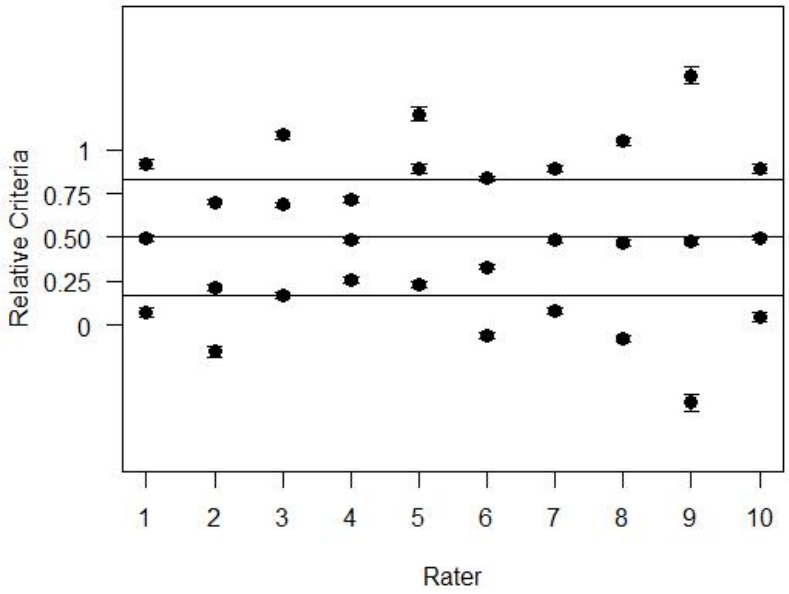


Figure 4.12. BIB Design, Criteria Parameters for a 4-class Ordered Perception Model with Rater Effects, Fit Equal Perception Models. The plot shows relative response criteria for each rater, SE bars, and optimal location points as lines.

Summary

This section shows that fitting equal perception models to ordered perception data leads to deviations in estimates, as expected. Fully-crossed design gave slightly better estimates than BIB design. Plots were useful to detect rater effects despite simulation design. Interestingly, even though a wrong model was fitted, rater effects were still detected. Therefore, if the main purpose of using SDT models is to detect rater effects, it may be alright to simply fit the parsimonious equal perception model yet look at he ordered results as a check.

This simulation can answer the second research question: To what extent will fitting wrong models affect parameter recovery? Results show that fitting equal perception models to ordered perception data leads to deviations in estimates, especially for BIB design. However, rater effects could always be determined and that is the important point.

4.4 Simulation 4: Equal Perception Data, Fit Ordered Perception Model

Without rater effects

Table A4.1 shows the parameter recovery of *d's* and *c's* over 100 replications of the fully-crossed design without rater effects. Recovery of *d*'s and *c*'s was excellent, with percent deviations all below 5%. The recovery of the latent class sizes was also excellent, with percent deviations all below 1%. Figure 4.13 shows no rater effects as all the circles overlapped with the triangles.

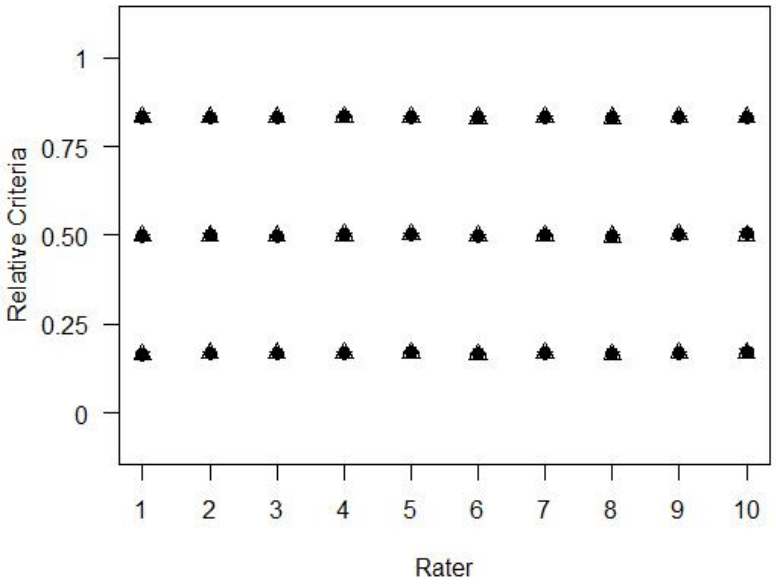


Figure 4.13. Fully Crossed Design, Criteria Parameters for a 4-class Equal Perception Model Without Rater Effects, Fit Ordered Perception Model. The plot shows relative response criteria for each rater and SE bars, with open triangles as predicted relative criteria from *d* estimates and filled circles as obtained relative criteria estimates.

Table B4.1 shows the parameter recovery of *d's* and *c's* over 100 replications of the BIB design without rater effects. Compared with its fully-crossed design counterpart, the quality of parameter recovery was a little worse but still reasonable. Most parameters were deviated by below 5%, with a few first *d*'s and *c*'s underestimated by 10%-30%. For estimates of latent class sizes, the first and fourth classes were overestimated by around 15%, whereas the middle ones

were underestimated by slightly over 5%. Same as in previous simulations, missing data to some extent degraded the estimation of some parameters, especially the first and fourth latent class sizes. Figure 4.14 shows that detection of rater effects was reasonable as most circles were close to the triangles.

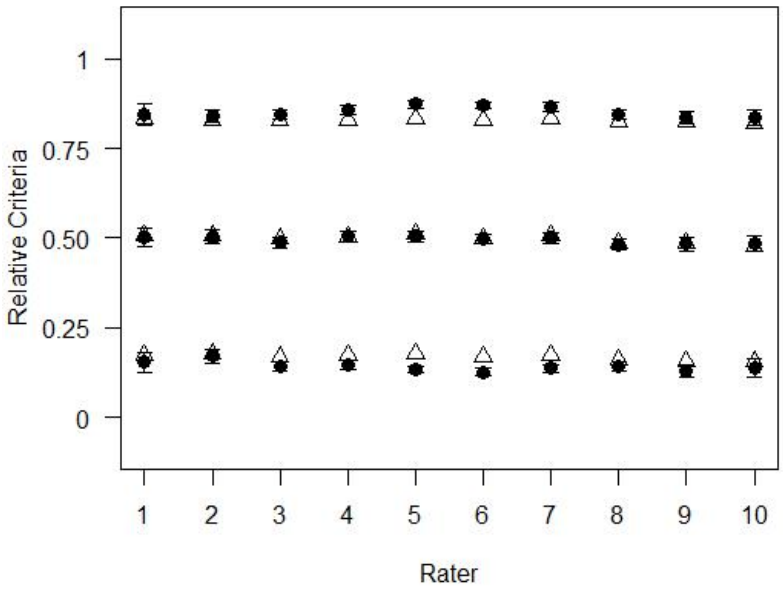


Figure 4.14. BIB Design, Criteria Parameters for a 4-class Equal Perception Model Without Rater Effects, Fit Ordered Perception Model. The plot shows relative response criteria for each rater, with open triangles as predicted relative criteria from *d* estimates and filled circles as obtained relative criteria estimates.

With rater effects

Table A4.2 shows the parameter recovery of *d's* and *c's* over 100 replications of the fully-crossed design with rater effects. Recovery of *d*'s and *c*'s was excellent, with percent deviations almost all below 5%. The recovery of the latent class sizes was also excellent, with percent deviations all below 2%. Figure 4.15 shows that all the rater effects were determined. Actually, if the triangles were changed to lines, this figure would look nearly the same as its counterpart Figure 4.3 Panel (B) where the same data were fitted with the equal perception

model, which means that detection of rater effects would not be affected if the ordered perception model is fitted to equal perception data.

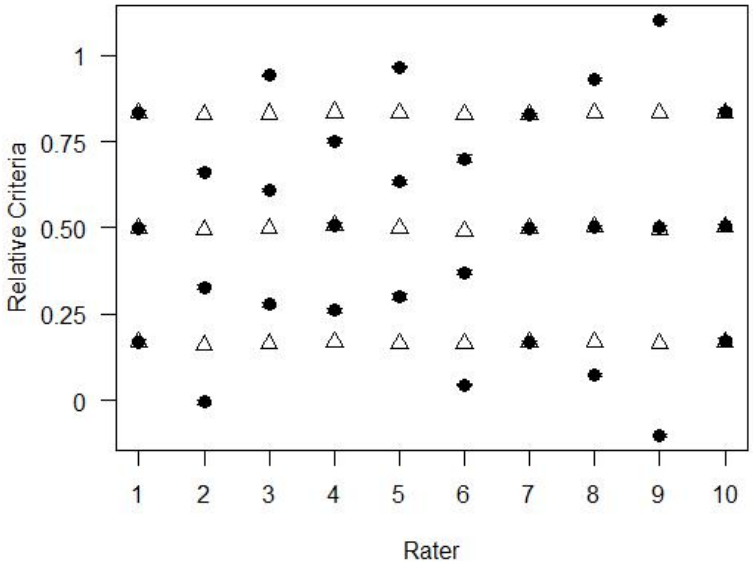


Figure 4.15. Fully Crossed Design, Criteria Parameters for a 4-class Equal Perception Model with Rater Effects, Fit Ordered Perception Model. The plot shows relative response criteria for each rater and SE bars, with open triangles as predicted relative criteria from *d* estimates and filled circles as obtained relative criteria estimates.

Table B4.2 shows the parameter recovery of *d's* and *c's* over 100 replications of the BIB design with rater effects. Compared with its fully-crossed design counterpart, the quality of parameter recovery was worse. Some *d*'s and *c*'s were largely deviated from the true value and most of these large deviations were around 20%. For estimates of latent class sizes, the first and fourth classes were overestimated by slightly over 15%, whereas the middle ones were underestimated by slightly over 5%. Again, missing data to some extent degraded the estimation of some parameters.

Figure 4.16 shows that it was possible to detect all rater effects by comparing the locations of SE bars to triangles. Compared with its counterpart Figures 4.4 Panel (B), the

optimal locations in Figure 4.16 were not on a line but this did not affect detection of rater effects.

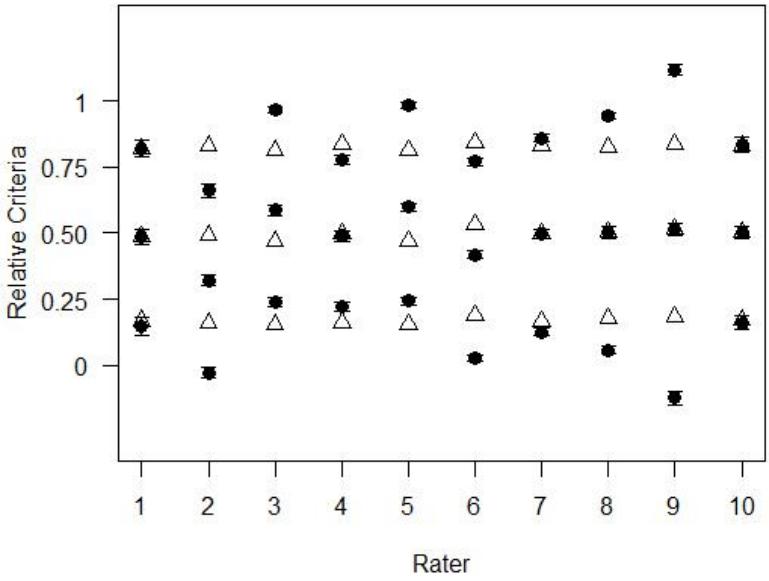


Figure 4.16. BIB Design, Criteria Parameters for a 4-class Equal Perception Model with Rater Effects, Fit Ordered Perception Model. The plot shows relative response criteria for each rater, with open triangles as predicted relative criteria from *d* estimates and filled circles as obtained relative criteria estimates.

Summary

This section shows that fitting ordered perception models to equal perception data leads to deviations in some parameter estimates in the BIB design, as expected. Fully-crossed design gave slightly better estimates. Also, data without rater effects tended to have slightly better parameter recovery than data with rater effects. Plots were useful to detect rater effects despite simulation design. Same as results in simulation 3, even if a wrong model was fitted, rater effects were still detected. However, if the main purpose of using SDT models is to detect rater effects, it may be reasonable to simply fit the parsimonious equal perception model yet check the results of the ordered model.

This simulation answers the second research question: To what extent will fitting wrong models affect parameter recovery? Results show that fitting ordered perception models to equal

perception data gave excellent parameter recovery except for BIB data with rater effects. Rater effects could always be determined.

4.5 Model Selection

Table 4.1 shows the proportion of times true model was recovered by the LR-like test and information criteria. For the LR test, the null hypothesis is that the equal and ordered perception models are equivalent. The difference of -2×Log-likelihood between these two models, which asymptotically follows the chi-square distribution, is compared to the critical value at p=0.05. If the difference of -2×Log-likelihood is bigger than the critical value, then the null hypothesis is rejected, with sufficient evidence to deny the equivalency of the two models. The ordered perception model is selected. Otherwise, the equivalency hypothesis cannot be rejected and the equal perception model is selected. Comparing results from fitting both equal and ordered perception models to the same data in the four simulation studies answers the third research question: How does model fit compare for the ordered and equal perception models?

Table 4.1. Performance of Fit Indices, N = 1,000

True Model	Experimental Condition		LR	AIC	BIC
	Data Design	Rater Effect		Type I Error	
Equal Perception	Fully-Crossed	No	0.07	0.01	0.00
		Yes	0.05	0.01	0.00
	BIB	No	0.05	0.00	0.00
		Yes	0.00	0.00	0.00
				Power	
Ordered Perception	Fully-Crossed	No	1.00	1.00	1.00
		Yes	1.00	1.00	1.00
	BIB	No	0.94	0.63	0.00
		Yes	0.54	0.23	0.00

Table 4.1 indicates that the LR-like test could always pick the right model. Both the top and bottom parts are showing the percent of times the equal model was rejected. The top shows

the Type I error of the LR test, where the true equal perception model was rejected less than 10% of times. Since the null hypothesis was 'equal fits', the bottom part of the table shows the power of the LR test, where the false equal perception model was rejected 54% to 100% of times when the ordered model was true.

For fully-crossed design, AIC and BIC always pick the correct model. For BIB design, the results were mixed. When the true model was equal perception, AIC and BIC could always pick the correct model. By contrast, when the true model was ordered perception with rater effects, AIC and BIC favored the wrong yet parsimonious model. AIC could determine the correct model when there were no rater effects. Therefore, LR appears to be useful for model selection.

4.6 Real World Analysis: Language Test

Table 4.2 shows the score frequencies and number of essays scored by each rater. The total number of essays scored by each rater ranges from 61 to 354, and the median is 155. The total number of score 5 assigned by each rater ranges from 7 to 88, and the median is 28. The total number of score 4 ranges from 13 to 88, and the median is 38 .The total number of score 3 ranges from 7 to 62, and the median is 27. The total number of score 2 ranges from 12 to 82, and the median is 40. The total number of score 1 ranges from 4 to 60, and the median is 22.

Discrimination parameter (*d*'s) estimates

Table C1 presents parameter estimates for fitting the ordered perception model to a large-scale language test where 27 raters assigned scores 1-5 to essays of 2,350 students. As shown in Table C1, some raters such as 6, 9, 10, 14, 15, 16, and 21 had large estimates of *d*'s, with first *d*'s around 5 and last *d*'s around 20, which means that these raters had excellent performance in terms of discriminating essays of different qualities.

Table 4.2. Score Frequencies and Number of Essays by Each Rater for Language Test Data

Rater	Score					Total
	5	4	2	3	1	
1	46	34	27	37	35	179
2	25	28	14	23	10	100
3	19	45	18	42	22	146
4	36	50	38	48	37	209
5	30	54	48	57	33	222
6	37	51	33	82	42	245
7	38	39	41	56	38	212
8	14	15	25	30	17	101
9	42	56	29	64	38	229
10	9	38	44	51	28	170
11	88	73	40	64	48	313
12	27	47	43	46	21	184
13	68	88	62	76	60	354
14	31	36	14	26	22	129
15	27	40	25	40	23	155
16	39	82	52	77	40	290
17	36	65	48	64	30	243
18	24	29	30	36	16	135
19	34	33	22	19	14	122
20	35	34	24	42	20	155
21	28	30	32	30	19	139
22	14	29	11	23	11	88
23	18	21	9	12	11	71
24	17	24	17	35	4	97
25	12	43	17	35	23	130
26	9	23	7	14	8	61
27	7	13	12	16	17	65

Figure 4.17 shows the d_{jm}'s for all the raters. Overall, equal spacing looks reasonable since the distances between adjacent perception distributions do not diverge for many raters. Obvious examples are Raters 5, 13, and 21. On the other hand, the plot shows that it is reasonable to have weaker assumptions of distances between adjacent perception distributions for other raters. For example, for Rater 2 the distance between latent categories of 4 and 5 is much smaller than those between other latent categories, meaning that it is difficult for Rater 2 to distinguish 4 from 5. For Rater 3, it is difficult to distinguish between latent categories of 3 and 4.

For Rater 8, it is difficult to distinguish between latent categories of 2 and 3. For Rater 26, it is more difficult to distinguish between latent categories at the end than between those in the middle. For Rater 23, it is more difficult to distinguish between latent categories of higher scores than between those of lower scores. In addition, the plot shows that Raters 6, 9, 10, 14, 15, 16, and 20 had the highest precision of rating since their *d* estimates were highest in the plot.

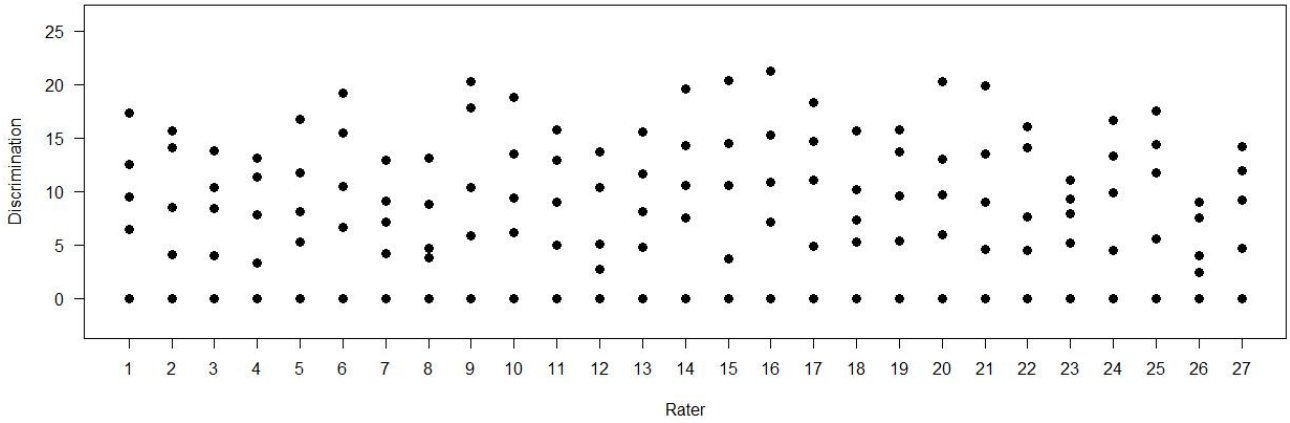
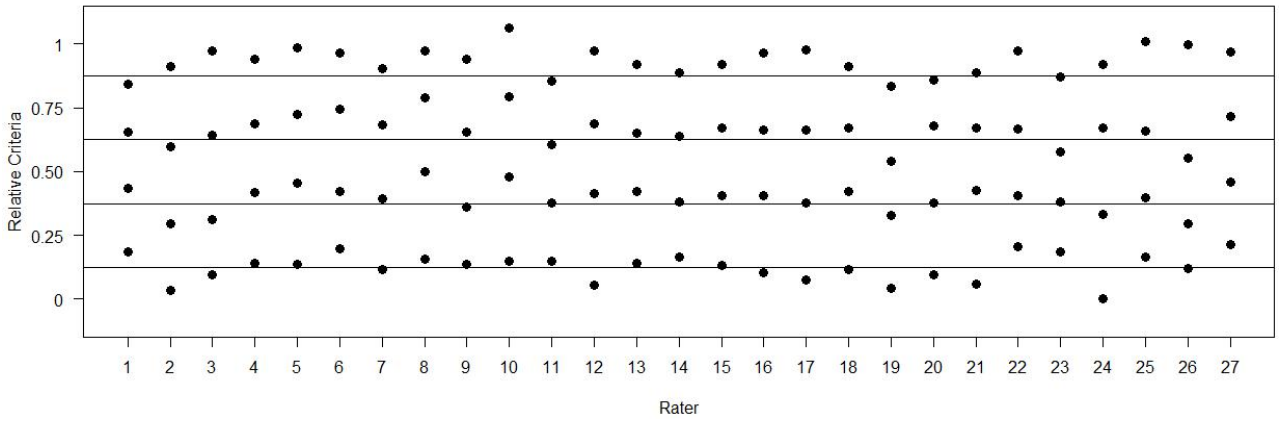


Figure 4.17. Distance Parameters for a 5-class Unequal Perception SDT Model, 27 Raters. The plot shows raw discrimination parameter estimates for each rater.

Criteria parameter (*c*'s) estimates

In Figure 4.18, a comparison of Panel A to Panel B shows that raters had similar effects for both equal and ordered perception models. For example, in Panel A, Rater 27 cleanly showed severity (shifted up) and 17 clearly showed centrality (top and bottom criterion shifted outwards). In Panel B, although the spacing changed, again 27 showed severity (shifted up) and 17 showed centrality. Also, in both Panel A and Panel B, Rater 1 showed extremity, Raters 10 and 27 were strict, and Rater 19 were lenient. On the contrary, many raters like 4, 7, 13, 14, 15, 18, and 21 did not show any obvious effects, and others like 24 had effects on specific scores.

(A)



(B)

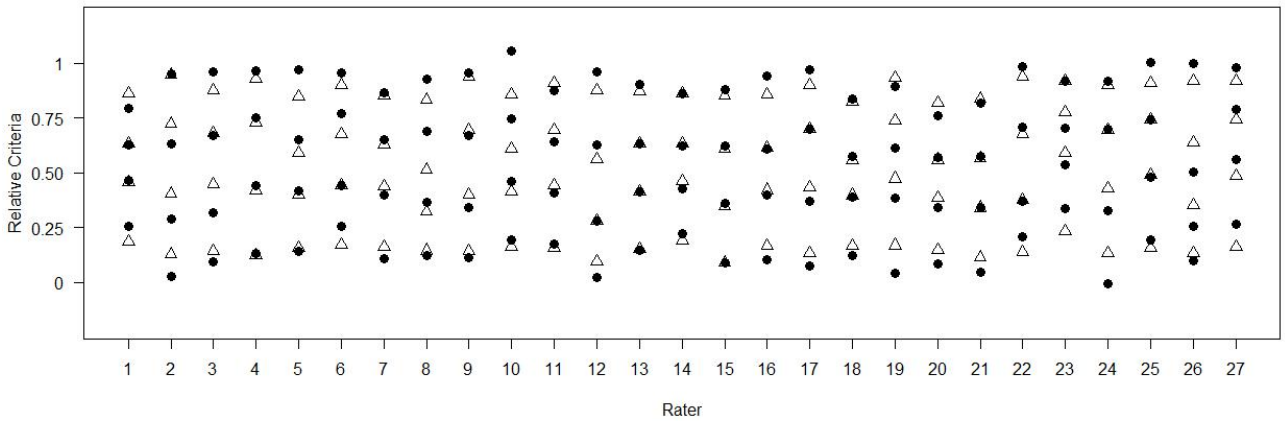


Figure 4.18. Criteria Parameters for 5-class SDT Models, 27 Raters. Panel (A) shows relative response criteria for each rater for equal perception model, with optimal locations as lines. Panel (B) shows relative response criteria for each rater for unequal perception model, with open triangles as predicted relative criteria from *d* estimates and filled circles as obtained relative criteria estimates.

Latent class sizes

Table 4.3 shows that the class sizes in the ordered perception model were similar to those of the fitted equal perception model. The SE's were small. The estimated sizes of the latent classes in both models did not follow an approximately normal distribution but appeared slightly skewed. Specifically, the class sizes of the lower two scores, slightly over 15%, were smaller

than those of the higher three scores, at around 20%-25%. The distribution of the quality of the essays in this test was skewed to the right.

Table 4.3. Estimated Sizes of Latent Classes for Equal and Ordered Perception Models for Language Test Data

Parameter	Equal Perception Model		Order Perception Model	
	Estimate	SE	Estimate	SE
Class 1	0.153	0.010	0.160	0.010
Class 2	0.136	0.013	0.165	0.015
Class 3	0.246	0.015	0.243	0.018
Class 4	0.240	0.016	0.229	0.020
Class 5	0.225	0.014	0.203	0.016

Model selection

Table 4.4 shows how fit indices selected between equal and ordered perception models. By comparing results from fitting these two models, the third research question can be answered: How does model fit compare for the ordered and equal perception models?

Table 4.4. Results for Equal and Ordered Perception Models for Language Test Data

Model	-2LL	AIC	BIC
Equal perception model	215658.9	215936.8586	216737.8004
Ordered perception model	215576.4	216016.4147	217284.0923
Difference=LR, *df*=86	82.5 (*p*=0.587)		

Table 4.4 shows that LR, AIC and BIC all favored the equal perception model. Figure 4.17 shows, for most raters, the distances between adjacent perception distributions tended to be equal. Thus, the equal SDT model might be adequate in the typical case.

4.7 Summary

This chapter has discussed results of analyzing simulated and real data and several conclusions can be made. First, the usefulness of plots to detect various rater effects was shown

in both simulation studies and real data analyses. Especially, a new plot was designed to show rater effects in ordered perception models, with triangles representing optimal criteria and supplanting the lines in the equal perception model. By comparing the locations of circles to triangles, it was easy to detect whether there are rater effects and what types.

Second, although parameter recovery deteriorated for BIB data, with rater effects, or for wrong models, the estimates were still reasonable and rater effects could always be determined. If the purpose is to detect rater effects, it may be alright to simply fit the parsimonious equal perception model yet look at he ordered results as a check.

Third, LR-like test may be the choice to pick the right model. As shown previously, LR-like tests had high power and low Type I error, rejecting the wrong null hypothesis or failing to reject the correct null hypothesis for at least more than half of the replications in each simulation condition. On the other hand, AIC and BIC favored parsimonious models for BIB data, especially when there were rater effects, even if the data were generated from the complex model. AIC might work well when there were no rater effects for BIB data although rater effects were unlikely to be nonexistent in reality. AIC and BIC could pick the correct model for fully-crossed data although it is rare for testing companies to use fully-crossed design.

Chapter 5

Summary and Discussion

5.1 Summary

The purpose of this study was to examine parameter recovery, model performance, and model fit for the ordered perception SDT rater model. To fulfil this purpose, four simulation studies were carried out for both fully-crossed and BIB designs, both without and with rater effects. Simulation 1 fitted correct models to equal perception data to provide a baseline for comparison with other simulation studies. Simulation 2 fitted correct models to ordered perception data to answer the first research question about parameter recovery for the ordered perception model. Simulation 3 fitted equal perception models to the data generated in simulation 2 to answer the second research question of model performance and the third research question of model fit. Simulation 4 fitted ordered perception models to the data generated in simulation 1 to answer the second and third research questions. Real data were analyzed to answer research Question 3.

Parameter recovery tended to be excellent for the ordered perception model. However, estimates in the BIB design were poorer. Data without rater effects tended to have slightly better parameter recovery than data with rater effects. Plots were useful to detect rater effects despite simulation design.

Fitting equal perception models to ordered perception data or vice versa tended to give different estimates of d_j, as expected, and c_j as well, even though the effects were generally small.

Most importantly, rater effects were still apparent. Fully-crossed design gave slightly better estimates than BIB design.

An interesting finding was that even though a wrong model was fitted, rater effects were still detected in the plot. The BIB data might create centrality effects for some raters. If the main purpose of using SDT models is to detect rater effects, it may be reasonable to always fit the equal perception model and then use the plot to show rater effects.

Using plots to show rater effects is simple, compared with statistical indexes such as the *outfit* statistic and the *infit* statistic which respectively indicate severity and centrality (Wolfe, Chiu, & Myford, 2000; Wright & Masters, 1982). The former is the mean of the squared standardized residuals between the observed response scores and the expected scores, and the latter is the variance-weighted mean of squared standardized residuals between the observed response scores and the expected scores. These statistics are designed for the FACETS model.

Plots are also more vivid than statistics of rater accuracy, such as rater agreement indices (von Eye &Mun, 2005), true ratings based on average ratings (Wolfe & McVay, 2012), distance between observed and true ratings based on expert judgment (Engelhard, 1996; 2013), and approaches based on the GT incorporating information from both group and individual levels (Marcoulides & Drezner, 1993; 1997; 2000).

Overall, compared with statistical measures, a graph seems more effective to show rater effects (DeCarlo et al., 2011). DeCarlo et al. illustrated such rater effects as severity, centrality, end effects, and preference of certain scores, all on a single graph. This study also showed all these effects in the plot for equal perception models, as well as the rater effects for the ordered perception model with a newly-designed plot. By comparing each rater's relative estimated

criteria with the relative optimal locations, it is easy to detect all the rater effects. A graph is especially appealing to those without training in statistics.

5.2 Practical Implications

This study explores rater effects in evaluation of CR items. With added flexibility to the discrimination parameter, the ordered perception model makes weaker assumptions. Using the information provided by the model, testing institutions can offer training to raters identified with rater effects so that the scores test takers receive are more objective. Meanwhile, reliable estimation of rater effects requires a sufficient number of CR items scored by each rater. It was found that if a rater scored fewer than 60 items for scores 1 to 5, estimates of *d*'s and *c*'s for this rater tended to be insignificant. For constructed responses of more than five categories, each rater should score more than 60 items to generate significant estimates. Considering the high cost of training raters, it may be advisable for testing companies to maintain long-term contracts with raters showing little effects.

Also, the ordered perceptional SDT model can help to improve the scoring rubric. If many raters have the same effects, then the problem may lie in the specifications of the rubric. For instance, if most raters do not assign the highest score, then the description of that score should be checked to see whether it is too harsh.

5.3 Limitations and Future Research

For one limitation, this study only attempts to simulate data with specific conditions. To generalize the results, more conditions may be tried in the future. For instance, CR items of more than four categories may be studied. Other missing data designs, such as unbalanced, incomplete block (UIB) or spiral design, are possible directions of study. The effects of small sample sizes on parameter estimation are also worth trying, where strong priors will probably be needed.

Another direction of study is comparing rater effects shown in different numbers of latent categories of CR items. For example, it is not known whether it is easier for raters to show specific effects, such as centrality, when there are three or four latent categories. Also, when the number of latent categories is bigger than, say, six, the number of parameters will be too large for the estimation of the ordered perception model to converge.

There have been little attempt to study the effects of estimating parameters in a model combing both MC items and covariates, so another possibility in the future is to incorporate MC items into the HRM-SDT model (Kim, 2009; Mariano & Junker, 2007). With regard to the relationship between the examinee proficiency θ and scores of MC items, polytomous IRT models may be used. In addition, both MC items and covariates can be included in the HRM-SDT model with a single CR item.

The condition where the number of perception distributions is different from that of scores in the scoring rubric may be studied. For instance, while the rubric specifies 4 scores, 3 perception distributions can be estimated. Then, how this may affect parameter recovery and detection of rater effects can be examined.

The ordered perception model is a type of ordered clustered model discussed by Croon (1990) and Vermunt and Magidson, (2016). In the situation with observed Y, equal d across the categories is proportional odds, same as in the equal perception model (DeCarlo, 1998). A test of proportional odds compares this against the situation where d varies across the categories, same for the unequal perception model with latent categories η. So comparing the two with, say, a likelihood ratio test is the same as a test of proportional odds (DeCarlo, 1998). LG was used to implement the parameter restriction (the ordered d's) and LG does it using Croon's (1990)

approach of restricting the probabilities, as mentioned in the technical manual (Vermunt & Magidson, 2016). The monotonic restriction was placed on the *d* parameters using LG.

However, the LR, AIC, and BIC indices used in the current study are not really the usual statistics based on the -2LL because the PME with Bayes constants of 1 rather than the MLE was used in model estimation. These indices were LR, AIC, and BIC computed on the -2LL reported in LG for PME. In the presence of PME, the reported LL's are more like posteriors which have been smoothed slightly by the priors on the response probabilities (Vermunt & Magidson, 2016). There is a huge related literature that we have noted, but it brings up issues in estimation, testing, and so on, that need closer attention. For example, there is a huge discussion about testing monotonicity (Vermunt, 2001; Vermunt & Magidson, 2016), so the LR examined here barely cracks the surface. Basically, the monotone *d*'s lead to stochastic ordering.

Since the goodness-of-fitness measures do not have asymptotic distributions with inequality constrains applied, Vermunt (2001) suggested use of parametric bootstrapping to obtain *p* values and compare models. However, more research is needed to study how effective bootstrapping is in evaluating model performance and how it compares with other indices such as the LR-like test, AIC, BIC, and DIC. In addition, cross-validation methods may be explored in terms of comparing equal and ordered perception models.

Also, the SDT ordered perception rater model is a type of restricted latent class models which put both equality and inequality restrictions on sums of conditional probabilities. The restrictions make the ordered perception model related to both parametric and nonparametric IRT models (Vermunt, 2001). For example, the monotonicity restriction on the logits of cumulative probabilities of adjacent latent classes makes the ordered perception model similar to the

polytomous IRT models, such as the graded response model (GRM) (Samejima, 1969). Both models are generalized linear models using cumulative category logits (Agresti, 2013; Dobson & Barnett, 2008). The GRM represents the difference of two cumulative latent categories on the logit scale. The main difference is that the GRM uses continuous latent classes yet the ordered perception model discrete latent classes. In a sense, the ordered perception model is a type of semi-parametric IRT polytomous model.

Another interesting area worthy of further study is to generalize the ordered perception model to multivariate situations, like analytic scoring. The current study is a univariate case where one essay of each test taker was analyzed. Yet, in reality some tests, such as GRE and TOEFL, have two essays, and other tests like SAT and ACT let each rater give four or five scores for each essay. These conditions need a multivariate ordered perception model to incorporate the different dimensions.

Another potential direction for future study is to simulate missing patterns from real data. For example, the missing pattern in the language test can be simulated 100 times with the LG-estimated parameter values to see how parameters will be recovered and whether the real rater effects can be determined with the newly designed plot in the current study.

A study having implications for testing companies is how many raters may provide satisfactory parameter recovery. Now, the common practice by testing companies is that two raters score the same essay and a third rater will score this essay if the scores by the first two raters diverge too much, e.g., 1 point in former SAT essay scoring. Results from the simulation studies in the BIB design show that parameter recovery degenerated compared with that of the fully-crossed design and that the true ordered perception model with rater effects could not be recovered by AIC or BIC. One reason may be that two raters cannot provide sufficient

information for LG to recover either the parameters or the true model. Further research may be done to find out whether three or more raters can improve the parameter and model recovery.

Another possibility for future research involves estimation methods. To improve estimation results where the MLE method would not converge for large *d*'s, the current study used the PME method for all the models. The PME algorithm converges fast but cannot give the distribution of the estimates. To tackle this, the Bayesian algorithms such as the MCMC approach may be used to estimate the parameters. Results may be compared with those of the PME algorithm.

Unequal scaling/variance may be assumed. However, unequal scaling means that the rater scores will not be ordered, so stochastic ordering of scores has to be used to make scoring ordered. Additionally, when scaling is unequal, the likelihood ratio of different categories will not be monotonic and appears to be counterintuitive. A random parameter *d* needs to be assumed and parameters *c*'s need to be constrained so that one *c* is always larger than another and will not cross (DeCarlo et al., 2011).

The Bayesian optimal criteria location may be explored. In medicine, most people do not have cancer so it seems appropriate to put the criteria location on the side of the distribution of patients with cancer. Similarly, to increase the accuracy of prediction, it seems proper to put the location criteria on the side of the distribution that has a smaller proportion or larger frequency.

Finally, cross-validation commonly used in predictive modeling may be attempted in comparison of the HRM SDT models. The charm of cross-validation is that the model is trained and tested with different datasets, so the model finally selected tends to avoid overfitting and have better predictive power.

References

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Hoboken, NJ: Wiley.

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Budapest: Akademiai Kiado.

Ando, T. (2011). Predictive Bayesian model selection. *American Journal of Mathematical and Management Sciences*, *31*(1-2), 13-38.

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., & Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association,92* (440), 1375-1386.

Bock, R. D., Brennan, R. L., & Muraki, E. (2002). The information in multiple ratings. *Applied psychological measurement*, *26*(4), 364-375.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Brennan, R. L. (1992). An NCME instructional module on generalizability theory. *Educational Measurement: Issues and Practice, 11*(4), 27–34.

Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer-Verlag.

Briggs, D. C., & Wilson, M. (2007). Generalizability in item response modeling. *Journal of Educational Measurement*, *44*(2), 131-155.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological methods & research*, *33*(2), 261-304.

Casabianca, J. M., & Junker, B. (2013). Hierarchical rater models for longitudinal assessments. *Annual Meeting of the National Council for Measurement in Education*, San Francisco, CA.

Casabianca, J. M., Junker, B. W., & Patz, R. (2012). The hierarchical rater model. *Handbook of modern item response theory. Boca Raton, FL: Chapman & Hall/CRC*.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, *90*(432), 1313-1321.

Chib, S., & Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association*, *96*(453), 270-281.

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging* (Vol. 330). Cambridge: Cambridge University Press.

Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 311–359). New York: Plenum.

Clogg, C. C., & Goodman, L. A. (1984). Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association, 79* (388), 762-771.

Croon, M. (1990). Latent class analysis with ordered latent classe. *British Journal of Mathematical and Statistical Psychology*, *43*(2), 171-192.

Dayton, C. M. (1998). *Latent class scaling analysis*. Thousand Oaks, CA: Sage.

Dayton, C. M., & Macready, G. B. (1988a). A latent class covariate model with applications to criterion-referenced testing. In R. Langeheine, & J. Rost (Eds.), *Latent trait and latent class models* (pp. 129–143). New York, NY: Plenum Press.

Dayton, C. M., & Macready, G. B. (1988b). Concomitant-variable latent-class models. *Journal of the American Statistical Association, 83* (401), 173-178.

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods*, *3*, 186–295.

DeCarlo, L. T. (2002). A latent class extension of signal detection theory, with applications. *Multivariate Behavioral Research*, *37*, 423–451.

DeCarlo, L. T. (2005). A model of rater behavior in essay grading based on signal detection theory. *Journal of Educational Measurement*, *42*, 53–76.

DeCarlo, L. T. (2008a, March). *On a hierarchical rater model for essay grading: incorporating a latent class signal detection model.* Paper presented at the Annual Meeting of the National Council on Measurement in Education. New York, NY.

DeCarlo, L. T. (2008b). *Studies of a latent class signal detection model for constructed response scoring.* ETS Technical Report, RR-08-63. Princeton NJ: Educational Testing Service.

DeCarlo, L. T., Kim, Y., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement*, *48*(3), 333-356.

DeCarlo, L. T., & Zhou, X. (2019). A signal detection model for rater scoring with ordered distributions. Manuscript in preparation.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1-38.

Dobson, A. J., & Barnett, A. (2008). *An introduction to generalized linear models*. CRC press.

Engelhard Jr, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*, 93-112.

Engelhard Jr, G. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 33*(1), 56-70.

Engelhard Jr, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge.

Ercikan, K., Schwarz, R. R., Julian, M. W., Burket, G. R., Weber, M. M. & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement, 35,* 137-154.

Fischer, G. H. (1973). The linear logistic model as an instrument in educational research. *Acta Psychologica, 37,* 359-374.

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48,* 3-26.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9* (4), 466-491.

Formann, A. K. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association, 87* (418), 476-486.

Galindo-Garre, F., & Vermunt, J. K. (2006). Avoiding boundary estimates in latent class analysis by Bayesian posterior mode estimation. *Behaviormetrika*, *33*, 43–59.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis*, third edition. Boca Raton, FL: CRC press.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel hierarchical models* (Vol. 1). New York, NY, USA: Cambridge University Press.

Gelman, A., & Meng, X. L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, 163-185.

Green, D. M., & Swets, J. A. (1988). *Signal detection theory and psychophysics* (Rev. Ed.). Los Altos, CA: Peninsula Publishing.

Hombo, C., and Donoghue, J. R. (2001). Applying the hierarchical raters model to NAEP. Paper presented at the *Annual Meeting of the National Council on Measurement in Education*, Seattle,Washington.

Huang, G. H., & Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika, 69* (1), 5-32.

Hung, L. F., & Wang, W. C. (2012). The generalized multilevel facets model for longitudinal data. *Journal of Educational and Behavioral Statistics*, *37*(2), 231-255.

Jin, K. Y., & Wang, W. C. (2017). Assessment of Differential Rater Functioning in Latent Classes with New Mixture Facets Models. *Multivariate Behavioral Research*, 1-12.

Johnson, M. S. (2012). Bayesian inference using Gibbs sampling (BUGS) for IRT models. Invited chapter for WJ van der Linden & RK Hambleton. *Handbook of Modern Item Response Theory*.

Johnson, V. E., & Albert, J. H. (2006). *Ordinal data modeling*. Springer Science & Business Media.

Junker, B. W., Patz, R. J., & VanHoudnos, N. (2012). Markov Chain Monte Carlo for Item Response models. Invited chapter for WJ van der Linden & RK Hambleton. *Handbook of Modern Item Response Theory*.

Kamakura, W. A., Wedel, M., & Agrawal, J. (1994). Concomitant variable latent class models for conjoint analysis. *International Journal of Research in Marketing, 11*, 451-464.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773-795.

Kim, Y. (2009). *Combining constructed response items and multiple choice items using a hierarchical rater model*. Columbia University.

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement*, *33*(7), 519-537.

Linacre, J. M. (1989). *Many-faceted Rasch Measurement.* Chicago, IL: MESA Press.

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed-response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement, 31,* 234-250.

Marcoulides, G. A., & Drezner, Z. (1993). A procedure for transforming points in multidimensional space to a two-dimensional representation. *Educational and Psychological Measurement, 53*, 933-940.

Marcoulides, G. A., & Drezner, Z. (1997). A method for analyzing performance assessments. In M. Wilson, G. Engelhard Jr., & K. Draney (Eds.), *Objective measurement: Theory into practice* (Vol. 4, pp. 261-277). Norwood, NJ: Ablex.

Marcoulides, G. A., & Drezner, Z. (2000). A procedure for detecting pattern clustering in measurement designs. In M. Wilson & G. Engelhard Jr (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 287-302). Norwood, NJ: Ablex.

Mariano, L. T. (2002). *Information accumulation, model selection and rater behavior in constructed response student assessments*. Unpublished doctoral dissertation, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA.

Mariano, L. T., & Junker, B.W. (2007). Covariates of the rating process in hierarchical models for multiple ratings of test items. *Journal of Educational and Behavioral Statistics*, *32*, 287–314.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*, 187–212.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

McLachlan, G., & Peel, D. (2000). *Finite mixture models.* New York: John Wiley & Sons.

Melton, B., Liang, K. Y., & Pulver, A. E. (1994). Extended latent class approach to the study of familial/sporadic forms of a disease: Its application to the study of the heterogeneity of schizophrenia. *Genetic Epidemiology, 11*, 311-327.

Muckle, T. J., & Karabatsos, G. (2009). Hierarchical generalized linear models for the analysis of judge ratings. *Journal of Educational Measurement*, *46*(2), 198-219.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4,* 386-422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5,* 189-227.

Neal, R. M. (2001). Annealed importance sampling. *Statistics and computing*, *11*(2), 125-139.

Patz, R. J. (1996). Markov Chain Monte Carlo methods for item response theory models with applications for NAEP. Ph.D. dissertation, Carnegie Mellon University, United States–Pennsylvania. Retrieved September 14, 2008, from Dissertations & Theses: Full Text database. (Publication No. AAT 9713184).

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, *27*(4), 341-384.

Pollack, J., Rock, D., & Jenkins, F. (1992, April). *Advantages and disadvantages of constructed-response items formats.* Paper presented at the Annual Meeting of the America Educational Research Association, San Francisco.

Rodriguez, M. C. (2002). Choosing an item format. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 213-231). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17*.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York, NY: Chapman &Hall.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461-464.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*(4), 298-321.

Spieglhalter, D. J., Best, N. G., Carlin, B. P., & Van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical, Series B*, *64*, 583-616.

Uebersax, J. S. (1993). Statistical modeling of expert ratings on medical treatment appropriateness. *Journal of the American Statistical Association, 88*(422), 421-427.

van Onna, M. J. H. (2004). Ordered latent class models in nonparametric item response theory. Unpublished doctoral dissertation, University of Groningen, Amsterdam, The Netherlands.

Verhelst, N. D., & Verstralen, H. H. (2001). An IRT model for multiple raters. In *Essays on item response theory* (pp. 89-108). Springer New York.

Vermunt, J. K. (2001). The use of restricted latent class models for defining and testing nonparametric and parametric item response theory models. *Applied Psychological Measurement, 25*(3), 283-294.

von Eye, A., & Mun, E. Y. (2005). Analyzing rater agreement: Manifest variable methods.

Mahwah, NJ: Erlbaum

Vermunt, J. K., & Magidson, J. (2016). Technical guide for Latent GOLD 5.1: basic, advanced, and syntax. Statistical Innovations. *Inc., Belmont Google Scholar*.

von Eye, A., & Mun, E. Y. (2005). Analyzing rater agreement: Manifest variable methods. Mahwah, NJ: Erlbaum

Wang, W. C., & Liu, C. Y. (2007). Formulation and application of the generalized multilevel facets model. *Educational and Psychological Measurement*, *67*(4), 583-605

Wang, W. C., & Wilson, M. (2005). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement*, *29*(4), 296-318.

Wang, Z. G. (2012). *On the use of covariates in a latent class signal detection model, with applications to constructed response scoring*. Columbia University.

Wesolowski, B. C., Wind, S. A., & Engelhard Jr, G. (2015). Rater fairness in music performance assessment: Evaluating model-data fit and differential rater functioning. *Musicae Scientiae*, *19*(2), 147-170.

Wesolowski, B. C., Wind, S. A., & Engelhard Jr, G. (2016). Rater analyses in music performance assessment: Application of the Many Facet Rasch Model. In *Connecting practice, measurement, and evaluation: Selected papers from the 5th International Symposium on Assessment in Music Education* (pp. 335-356).

Wolfe, E. W., Chiu, C. W., & Myford, C. M. (2000). Detecting rater effects in simulated data with a multi-faceted Rasch rating scale model. *Objective measurement: Theory into practice*, *5*, 147-164.

Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice*, *31*(3), 31-37.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*(p. 1982). Chicago: Mesa Press.

Wu, C. J. (1983). On the convergence properties of the EM algorithm. *The Annals of statistics*, 95-103.

Wu, S. M., & Tan, S. (2016). Managing rater effects through the use of FACETS analysis: the case of a university placement test. *Higher Education Research & Development*, *35*(2), 380-394.

Yamaguchi, K. (2000). Multinomial logit latent-class regression models: An analysis of the predictors of gender-role attitudes among Japanese women. *American Journal of Sociology, 105* (6), 1702-1740.

Zhu, X., & Stone, C. A. (2011). Assessing fit of unidimensional graded response models using Bayesian methods. *Journal of Educational Measurement*, *48*(1), 81-97.

Appendix A

Parameter Estimates, Bias/Deviation, Percent Bias/Deviation, and Mean Squared Error for the Latent Class SDT Model, Fully-Crossed Design

Table A1.1

Equal Perception Model Without Rater Effects, Fit Equal Perception Model, N = 1,000

Parameter Estimates					
Parameter	Value	Estimate	Bias	%Bias	MSE
d_1	1.00	0.993	-0.007	0.690	0.005
d_2	2.00	2.005	0.005	0.250	0.008
d_3	3.00	3.013	0.013	0.439	0.017
d_4	4.00	4.001	0.001	0.021	0.020
d_5	5.00	4.977	-0.023	0.455	0.032
d_6	5.50	5.476	-0.024	0.443	0.033
d_7	4.50	4.506	0.006	0.144	0.032
d_8	3.50	3.494	-0.006	0.157	0.016
d_9	2.50	2.499	-0.001	0.044	0.010
d_{10}	1.50	1.504	0.004	0.269	0.007
c_{11}	0.50	0.489	-0.011	2.176	0.014
c_{12}	1.50	1.483	-0.017	1.156	0.017
c_{13}	2.50	2.489	-0.011	0.455	0.020
c_{21}	1.00	0.992	-0.008	0.790	0.018
c_{22}	3.00	3.006	0.006	0.191	0.024
c_{23}	5.00	5.002	0.002	0.049	0.042
c_{31}	1.50	1.494	-0.006	0.388	0.027
c_{32}	4.50	4.509	0.009	0.210	0.043
c_{33}	7.50	7.519	0.019	0.252	0.104
c_{41}	2.00	1.980	-0.020	0.991	0.033
c_{42}	6.00	6.003	0.003	0.042	0.053
c_{43}	10.00	10.015	0.015	0.145	0.137
c_{51}	2.50	2.482	-0.018	0.720	0.042
c_{52}	7.50	7.463	-0.037	0.492	0.077
c_{53}	12.50	12.476	-0.024	0.192	0.206
c_{61}	2.75	2.715	-0.035	1.267	0.049
c_{62}	8.25	8.204	-0.046	0.562	0.118
c_{63}	13.75	13.693	-0.057	0.416	0.209
c_{71}	2.25	2.246	-0.004	0.181	0.040
c_{72}	6.75	6.748	-0.002	0.033	0.102
c_{73}	11.25	11.260	0.010	0.093	0.187
c_{81}	1.75	1.754	0.004	0.247	0.029
c_{82}	5.25	5.239	-0.011	0.204	0.037
c_{83}	8.75	8.736	-0.014	0.164	0.071

c_{91}	1.25	1.244	-0.006	0.468	0.021
c_{92}	3.75	3.738	-0.012	0.329	0.029
c_{93}	6.25	6.245	-0.005	0.084	0.054
c_{101}	0.75	0.772	0.022	2.953	0.014
c_{102}	2.25	2.275	0.025	1.109	0.019
c_{103}	3.75	3.754	0.004	0.107	0.028
		Latent Class Size			
Parameter	Value	Estimate	Bias	%Bias	MSE
Class 1	0.15	0.151	0.001	0.680	<0.001
Class 2	0.35	0.349	-0.001	0.225	<0.001
Class 3	0.35	0.350	<0.001	0.092	<0.001
Class 4	0.15	0.150	<0.001	0.059	<0.001

Table A1.2

Equal Perception Model With Rater Effects, Fit Equal Perception Model

		Parameter Estimates			
Parameter	Value	Estimate	Bias	%Bias	MSE
d_1	1.00	1.000	<0.001	0.023	0.007
d_2	2.00	1.989	-0.011	0.538	0.009
d_3	3.00	3.017	0.017	0.552	0.016
d_4	4.00	3.999	-0.001	0.030	0.027
d_5	5.00	5.000	<0.001	0.002	0.120
d_6	5.50	5.522	0.022	0.408	0.139
d_7	4.50	4.502	0.002	0.034	0.034
d_8	3.50	3.496	-0.004	0.117	0.017
d_9	2.50	2.496	-0.004	0.145	0.015
d_{10}	1.50	1.493	-0.007	0.469	0.007
c_{11}	0.50	0.492	-0.008	1.623	0.019
c_{12}	1.50	1.499	-0.001	0.038	0.018
c_{13}	2.50	2.502	0.002	0.097	0.022
c_{21}	0.00	-0.004	-0.004	-	0.014
c_{22}	2.00	1.992	-0.008	0.412	0.018
c_{23}	4.00	3.980	-0.020	0.490	0.033
c_{31}	2.50	2.519	0.019	0.747	0.029
c_{32}	5.50	5.541	0.041	0.738	0.065
c_{33}	8.50	8.541	0.041	0.482	0.127
c_{41}	3.00	3.021	0.021	0.690	0.042
c_{42}	6.00	6.017	0.017	0.290	0.086
c_{43}	9.00	8.984	-0.016	0.175	0.115
c_{51}	4.50	4.493	-0.007	0.145	0.137
c_{52}	9.50	9.514	0.014	0.148	0.496
c_{53}	14.50	14.510	0.010	0.067	1.074
c_{61}	0.75	0.753	0.003	0.442	0.035
c_{62}	6.25	6.281	0.031	0.499	0.168

c_{63}	11.75	11.815	0.065	0.550	0.562
c_{71}	2.25	2.253	0.003	0.112	0.043
c_{72}	6.75	6.763	0.013	0.197	0.095
c_{73}	11.25	11.245	-0.005	0.042	0.193
c_{81}	0.75	0.770	0.020	2.680	0.029
c_{82}	5.25	5.244	-0.006	0.119	0.049
c_{83}	9.75	9.746	-0.004	0.045	0.139
c_{91}	-0.75	-0.753	-0.003	0.362	0.021
c_{92}	3.75	3.770	0.020	0.530	0.043
c_{93}	8.25	8.267	0.017	0.211	0.123
c_{101}	0.75	0.752	0.002	0.249	0.022
c_{102}	2.25	2.244	-0.006	0.268	0.021
c_{103}	3.75	3.743	-0.007	0.178	0.030
		Latent Class Size			
Parameter	Value	Estimate	Bias	%Bias	MSE
Class 1	0.15	0.149	-0.001	0.694	<0.001
Class 2	0.35	0.348	-0.002	0.516	<0.001
Class 3	0.35	0.350	<0.001	0.108	<0.001
Class 4	0.15	0.152	0.002	1.644	<0.001

Table A2.1

Ordered Perception Model Without Rater Effects, Fit Ordered Perception Model

		Parameter Estimates			
Parameter	Value	Estimate	Bias	%Bias	MSE
d_{12}	1.00	1.008	0.008	0.755	0.038
d_{13}	2.00	2.018	0.018	0.876	0.035
d_{14}	3.00	3.018	0.018	0.613	0.070
d_{22}	1.00	1.006	0.006	0.624	0.041
d_{23}	3.00	3.032	0.032	1.051	0.051
d_{24}	5.00	5.035	0.035	0.697	0.087
d_{32}	3.00	3.028	0.028	0.926	0.090
d_{33}	6.00	6.033	0.033	0.551	0.096
d_{34}	7.00	7.017	0.017	0.248	0.153
d_{42}	1.00	1.038	0.038	3.843	0.054
d_{43}	5.00	5.009	0.009	0.188	0.072
d_{44}	6.00	6.013	0.013	0.222	0.110
d_{52}	5.00	5.045	0.045	0.900	0.204
d_{53}	6.00	6.022	0.022	0.368	0.189
d_{54}	11.00	11.114	0.114	1.037	0.348
d_{62}	5.50	5.540	0.040	0.732	0.329
d_{63}	11.00	11.044	0.044	0.396	0.468
d_{64}	16.50	16.541	0.041	0.247	0.728
d_{72}	1.00	1.022	0.022	2.156	0.042
d_{73}	5.50	5.526	0.026	0.464	0.110

d_{74}	6.50	6.503	0.003	0.043	0.137
d_{82}	1.00	0.990	-0.010	1.018	0.035
d_{83}	4.50	4.524	0.024	0.532	0.060
d_{84}	8.00	8.040	0.040	0.505	0.137
d_{92}	2.50	2.539	0.039	1.573	0.067
d_{93}	3.50	3.549	0.049	1.388	0.066
d_{94}	6.00	6.057	0.057	0.949	0.100
d_{102}	1.50	1.486	-0.014	0.952	0.049
d_{103}	3.00	2.991	-0.009	0.295	0.045
d_{104}	4.00	3.993	-0.007	0.171	0.074
c_{11}	0.50	0.516	0.016	3.114	0.025
c_{12}	1.50	1.517	0.017	1.144	0.026
c_{13}	2.50	2.511	0.011	0.438	0.030
c_{21}	0.50	0.504	0.004	0.826	0.030
c_{22}	2.00	2.009	0.009	0.447	0.037
c_{23}	4.00	4.033	0.033	0.833	0.052
c_{31}	1.50	1.493	-0.007	0.465	0.059
c_{32}	4.50	4.531	0.031	0.690	0.097
c_{33}	6.50	6.519	0.019	0.288	0.103
c_{41}	0.50	0.534	0.034	6.802	0.034
c_{42}	3.00	3.034	0.034	1.124	0.052
c_{43}	5.50	5.520	0.020	0.358	0.076
c_{51}	2.50	2.537	0.037	1.464	0.180
c_{52}	5.50	5.548	0.048	0.867	0.196
c_{53}	8.50	8.550	0.050	0.583	0.212
c_{61}	2.75	2.739	-0.011	0.385	0.179
c_{62}	8.25	8.309	0.059	0.711	0.367
c_{63}	13.75	13.824	0.074	0.541	0.535
c_{71}	0.50	0.506	0.006	1.248	0.029
c_{72}	3.25	3.264	0.014	0.418	0.066
c_{73}	6.00	6.007	0.007	0.121	0.099
c_{81}	0.50	0.514	0.014	2.880	0.028
c_{82}	2.75	2.741	-0.009	0.325	0.038
c_{83}	6.25	6.266	0.016	0.249	0.078
c_{91}	1.25	1.284	0.034	2.752	0.049
c_{92}	3.00	3.037	0.037	1.223	0.060
c_{93}	4.75	4.772	0.022	0.470	0.069
c_{101}	0.75	0.736	-0.014	1.801	0.030
c_{102}	2.25	2.239	-0.011	0.502	0.038
c_{103}	3.50	3.495	-0.005	0.143	0.046

Latent Class Size

Parameter	Value	Estimate	Bias	%Bias	MSE
Class 1	0.15	0.148	-0.002	1.074	<0.001
Class 2	0.35	0.352	0.002	0.602	<0.001
Class 3	0.35	0.349	-0.001	0.380	<0.001
Class 4	0.15	0.151	0.001	0.555	<0.001

Table A2.2

Ordered Perception Model With Rater Effects, Fit Ordered Perception Model

Parameter Estimates					
Parameter	Value	Estimate	Bias	%Bias	MSE
d_{12}	1.00	1.023	0.023	2.253	0.040
d_{13}	2.00	2.034	0.034	1.696	0.042
d_{14}	3.00	3.044	0.044	1.475	0.058
d_{22}	1.00	0.969	-0.031	3.114	0.035
d_{23}	3.00	2.960	-0.040	1.339	0.034
d_{24}	5.00	4.978	-0.022	0.450	0.099
d_{32}	3.00	3.079	0.079	2.624	0.150
d_{33}	6.00	6.110	0.110	1.839	0.168
d_{34}	7.00	7.086	0.086	1.232	0.169
d_{42}	1.00	1.007	0.007	0.746	0.084
d_{43}	5.00	4.985	-0.015	0.299	0.095
d_{44}	6.00	6.000	<0.001	0.002	0.171
d_{52}	5.00	5.073	0.073	1.466	1.034
d_{53}	6.00	6.049	0.049	0.824	1.054
d_{54}	11.00	11.124	0.124	1.123	1.356
d_{62}	5.50	5.556	0.056	1.014	0.643
d_{63}	11.00	11.209	0.209	1.897	1.256
d_{64}	16.50	16.900	0.400	2.425	2.486
d_{72}	1.00	1.050	0.050	4.964	0.054
d_{73}	5.50	5.572	0.072	1.309	0.095
d_{74}	6.50	6.587	0.087	1.333	0.140
d_{82}	1.00	1.019	0.019	1.869	0.051
d_{83}	4.50	4.530	0.030	0.671	0.080
d_{84}	8.00	8.004	0.004	0.049	0.195
d_{92}	2.50	2.490	-0.010	0.392	0.056
d_{93}	3.50	3.474	-0.026	0.730	0.067
d_{94}	6.00	6.019	0.019	0.316	0.104
d_{102}	1.50	1.532	0.032	2.144	0.038
d_{103}	3.00	3.033	0.033	1.106	0.047
d_{104}	4.00	4.022	0.022	0.548	0.064
c_{11}	0.50	0.514	0.014	2.711	0.037
c_{12}	1.50	1.533	0.033	2.221	0.034
c_{13}	2.50	2.539	0.039	1.570	0.037
c_{21}	-0.50	-0.539	-0.039	7.820	0.030
c_{22}	1.00	0.967	-0.033	3.347	0.023
c_{23}	3.00	2.959	-0.041	1.375	0.037
c_{31}	2.50	2.580	0.080	3.186	0.134
c_{32}	5.50	5.586	0.086	1.566	0.147
c_{33}	7.50	7.605	0.105	1.397	0.169

c_{41}	1.50	1.500	<0.001	0.007	0.063
c_{42}	3.00	2.991	-0.009	0.304	0.074
c_{43}	4.50	4.484	-0.016	0.366	0.083
c_{51}	4.50	4.554	0.054	1.199	1.023
c_{52}	7.50	7.546	0.046	0.608	1.087
c_{53}	10.50	10.645	0.145	1.381	1.346
c_{61}	0.75	0.730	-0.020	2.688	0.046
c_{62}	6.25	6.303	0.053	0.844	0.650
c_{63}	11.75	11.947	0.197	1.675	1.258
c_{71}	0.50	0.516	0.016	3.158	0.035
c_{72}	3.25	3.291	0.041	1.267	0.072
c_{73}	6.00	6.079	0.079	1.313	0.102
c_{81}	-0.50	-0.499	0.001	0.226	0.034
c_{82}	2.75	2.762	0.012	0.443	0.060
c_{83}	7.25	7.262	0.012	0.167	0.145
c_{91}	-0.75	-0.784	-0.034	4.521	0.026
c_{92}	3.00	2.965	-0.035	1.157	0.057
c_{93}	6.75	6.780	0.030	0.445	0.095
c_{101}	0.75	0.773	0.023	3.090	0.033
c_{102}	2.25	2.271	0.021	0.935	0.037
c_{103}	3.50	3.530	0.030	0.871	0.038

Latent Class Size

Parameter	Value	Estimate	Bias	%Bias	MSE
Class 1	0.15	0.152	0.002	1.512	<0.001
Class 2	0.35	0.349	-0.001	0.217	<0.001
Class 3	0.35	0.349	-0.001	0.201	<0.001
Class 4	0.15	0.149	-0.001	0.539	<0.001

Table A3.1

Ordered Perception Model Without Rater Effects, Fit Equal Perception Model

Parameter Estimates				
Parameter	Value	Estimate	Deviation	%Deviation
d_{12}	1.00	1.008	0.008	0.807
d_{13}	2.00	2.016	0.016	0.807
d_{14}	3.00	3.024	0.024	0.807
d_{22}	1.00	1.745	0.745	74.524
d_{23}	3.00	3.490	0.490	16.349
d_{24}	5.00	5.236	0.236	4.714
d_{32}	3.00	2.379	-0.621	20.696
d_{33}	6.00	4.758	-1.242	20.696
d_{34}	7.00	7.137	0.137	1.962
d_{42}	1.00	2.231	1.231	123.078
d_{43}	5.00	4.462	-0.538	10.769
d_{44}	6.00	6.692	0.692	11.539

d_{52}	5.00	2.516	-2.484	49.687
d_{53}	6.00	5.031	-0.969	16.145
d_{54}	11.00	7.547	-3.453	31.392
d_{62}	5.50	5.548	0.048	0.881
d_{63}	11.00	11.097	0.097	0.881
d_{64}	16.50	16.645	0.145	0.881
d_{72}	1.00	2.358	1.358	135.769
d_{73}	5.50	4.715	-0.785	14.266
d_{74}	6.50	7.073	0.573	8.816
d_{82}	1.00	2.711	1.711	171.132
d_{83}	4.50	5.423	0.923	20.503
d_{84}	8.00	8.134	0.134	1.675
d_{92}	2.50	1.707	-0.793	31.737
d_{93}	3.50	3.413	-0.087	2.482
d_{94}	6.00	5.120	-0.880	14.672
d_{102}	1.50	1.368	-0.132	8.816
d_{103}	3.00	2.736	-0.264	8.816
d_{104}	4.00	4.103	0.103	2.582
c_{11}	0.50	0.519	0.019	3.772
c_{12}	1.50	1.519	0.019	1.284
c_{13}	2.50	2.512	0.012	0.488
c_{21}	0.50	1.009	0.509	101.758
c_{22}	2.00	2.520	0.520	26.013
c_{23}	4.00	4.467	0.467	11.670
c_{31}	1.50	0.989	-0.511	34.081
c_{32}	4.50	3.704	-0.796	17.696
c_{33}	6.50	5.681	-0.819	12.594
c_{41}	0.50	1.236	0.736	147.136
c_{42}	3.00	3.364	0.364	12.117
c_{43}	5.50	5.490	-0.010	0.175
c_{51}	2.50	1.041	-1.459	58.373
c_{52}	5.50	3.786	-1.714	31.165
c_{53}	8.50	6.511	-1.989	23.396
c_{61}	2.75	2.783	0.033	1.212
c_{62}	8.25	8.335	0.085	1.035
c_{63}	13.75	13.930	0.180	1.311
c_{71}	0.50	1.302	0.802	160.466
c_{72}	3.25	3.542	0.292	8.976
c_{73}	6.00	5.766	-0.234	3.900
c_{81}	0.50	1.739	1.239	247.879
c_{82}	2.75	3.925	1.175	42.739
c_{83}	6.25	6.982	0.732	11.717
c_{91}	1.25	0.820	-0.430	34.413
c_{92}	3.00	2.556	-0.444	14.802
c_{93}	4.75	4.282	-0.468	9.847
c_{101}	0.75	0.627	-0.123	16.413

c_{102}	2.25	2.107	-0.143	6.365
c_{103}	3.50	3.366	-0.134	3.824
Latent Class Size				
Parameter	Value	Estimate	Deviation	%Deviation
Class 1	0.15	0.147	-0.003	1.989
Class 2	0.35	0.354	0.004	1.076
Class 3	0.35	0.348	-0.002	0.675
Class 4	0.15	0.152	0.002	1.054

Table A3.2

Ordered Perception Model With Rater Effects, Fit Equal Perception Model

Parameter Estimates				
Parameter	Value	Estimate	Deviation	%Deviation
d_{12}	1.00	1.011	0.011	1.088
d_{13}	2.00	2.022	0.022	1.088
d_{14}	3.00	3.033	0.033	1.088
d_{22}	1.00	1.664	0.664	66.355
d_{23}	3.00	3.327	0.327	10.903
d_{24}	5.00	4.991	-0.009	0.187
d_{32}	3.00	2.263	-0.737	24.570
d_{33}	6.00	4.526	-1.474	24.570
d_{34}	7.00	6.789	-0.211	3.019
d_{42}	1.00	2.641	1.641	164.081
d_{43}	5.00	5.282	0.282	5.632
d_{44}	6.00	7.922	1.922	32.040
d_{52}	5.00	2.238	-2.762	55.238
d_{53}	6.00	4.476	-1.524	25.397
d_{54}	11.00	6.714	-4.286	38.961
d_{62}	5.50	5.380	-0.120	2.187
d_{63}	11.00	10.759	-0.241	2.187
d_{64}	16.50	16.139	-0.361	2.187
d_{72}	1.00	2.426	1.426	142.645
d_{73}	5.50	4.853	-0.647	11.766
d_{74}	6.50	7.279	0.779	11.990
d_{82}	1.00	2.673	1.673	167.325
d_{83}	4.50	5.347	0.847	18.811
d_{84}	8.00	8.020	0.020	0.247
d_{92}	2.50	1.630	-0.870	34.796
d_{93}	3.50	3.260	-0.240	6.851
d_{94}	6.00	4.890	-1.110	18.495
d_{102}	1.50	1.372	-0.128	8.557
d_{103}	3.00	2.743	-0.257	8.557
d_{104}	4.00	4.115	0.115	2.874
c_{11}	0.50	0.501	0.001	0.177

c_{12}	1.50	1.519	0.019	1.290
c_{13}	2.50	2.525	0.025	0.986
c_{21}	-0.50	-0.155	0.345	69.070
c_{22}	1.00	1.407	0.407	40.724
c_{23}	3.00	3.349	0.349	11.633
c_{31}	2.50	1.747	-0.753	30.111
c_{32}	5.50	4.411	-1.089	19.795
c_{33}	7.50	6.599	-0.901	12.018
c_{41}	1.50	2.673	1.173	78.174
c_{42}	3.00	3.944	0.944	31.453
c_{43}	4.50	5.216	0.716	15.906
c_{51}	4.50	2.210	-2.290	50.887
c_{52}	7.50	5.240	-2.260	30.137
c_{53}	10.50	6.944	-3.556	33.865
c_{61}	0.75	0.704	-0.046	6.148
c_{62}	6.25	6.096	-0.154	2.459
c_{63}	11.75	11.498	-0.252	2.144
c_{71}	0.50	1.326	0.826	165.112
c_{72}	3.25	3.622	0.372	11.438
c_{73}	6.00	5.913	-0.087	1.443
c_{81}	-0.50	0.336	0.836	167.229
c_{82}	2.75	3.860	1.110	40.372
c_{83}	7.25	7.652	0.402	5.547
c_{91}	-0.75	-1.073	-0.323	43.004
c_{92}	3.00	2.419	-0.581	19.376
c_{93}	6.75	5.951	-0.799	11.835
c_{101}	0.75	0.625	-0.125	16.708
c_{102}	2.25	2.099	-0.151	6.710
c_{103}	3.50	3.363	-0.137	3.915

Latent Class Size

Parameter	Value	Estimate	Deviation	%Deviation
Class 1	0.15	0.154	0.004	2.970
Class 2	0.35	0.348	-0.002	0.566
Class 3	0.35	0.349	-0.001	0.320
Class 4	0.15	0.149	-0.001	0.903

Table A4.1

Equal Perception Model Without Rater Effects, Fit Ordered Perception Model

Parameter Estimates

Parameter	Value	Estimate	Deviation	%Deviation
d_{12}	1.00	0.986	-0.014	1.421
d_{13}	2.00	1.983	-0.017	0.829
d_{14}	3.00	2.981	-0.019	0.632
d_{22}	2.00	2.023	0.023	1.125

d_{23}	4.00	4.019	0.019	0.487
d_{24}	6.00	6.036	0.036	0.597
d_{32}	3.00	3.034	0.034	1.140
d_{33}	6.00	6.043	0.043	0.713
d_{34}	9.00	9.066	0.066	0.736
d_{42}	4.00	4.029	0.029	0.725
d_{43}	8.00	8.040	0.040	0.495
d_{44}	12.00	12.037	0.037	0.311
d_{52}	5.00	5.074	0.074	1.486
d_{53}	10.00	10.019	0.019	0.191
d_{54}	15.00	15.042	0.042	0.278
d_{62}	5.50	5.494	-0.006	0.100
d_{63}	11.00	10.971	-0.029	0.265
d_{64}	16.50	16.528	0.028	0.169
d_{72}	4.50	4.538	0.038	0.834
d_{73}	9.00	9.051	0.051	0.571
d_{74}	13.50	13.568	0.068	0.504
d_{82}	3.50	3.479	-0.021	0.612
d_{83}	7.00	6.980	-0.020	0.284
d_{84}	10.50	10.520	0.020	0.192
d_{92}	2.50	2.536	0.036	1.451
d_{93}	5.00	5.026	0.026	0.517
d_{94}	7.50	7.526	0.026	0.352
d_{102}	1.50	1.519	0.019	1.295
d_{103}	3.00	3.016	0.016	0.541
d_{104}	4.50	4.529	0.029	0.638
c_{11}	0.50	0.485	-0.015	3.079
c_{12}	1.50	1.479	-0.021	1.382
c_{13}	2.50	2.487	-0.013	0.535
c_{21}	1.00	1.003	0.003	0.312
c_{22}	3.00	3.019	0.019	0.636
c_{23}	5.00	5.017	0.017	0.335
c_{31}	1.50	1.509	0.009	0.628
c_{32}	4.50	4.528	0.028	0.629
c_{33}	7.50	7.539	0.039	0.521
c_{41}	2.00	2.002	0.002	0.099
c_{42}	6.00	6.036	0.036	0.593
c_{43}	10.00	10.049	0.049	0.488
c_{51}	2.50	2.553	0.053	2.107
c_{52}	7.50	7.545	0.045	0.606
c_{53}	12.50	12.549	0.049	0.394
c_{61}	2.75	2.735	-0.015	0.548
c_{62}	8.25	8.222	-0.028	0.338
c_{63}	13.75	13.728	-0.022	0.158
c_{71}	2.25	2.271	0.021	0.948
c_{72}	6.75	6.784	0.034	0.498

c_{73}	11.25	11.302	0.052	0.460
c_{81}	1.75	1.745	-0.005	0.302
c_{82}	5.25	5.229	-0.021	0.397
c_{83}	8.75	8.737	-0.013	0.151
c_{91}	1.25	1.270	0.020	1.615
c_{92}	3.75	3.769	0.019	0.500
c_{93}	6.25	6.274	0.024	0.390
c_{101}	0.75	0.782	0.032	4.249
c_{102}	2.25	2.286	0.036	1.613
c_{103}	3.75	3.767	0.017	0.441
Latent Class Size				
Parameter	Value	Estimate	Deviation	%Deviation
Class 1	0.15	0.151	0.001	0.663
Class 2	0.35	0.349	-0.001	0.225
Class 3	0.35	0.350	<0.001	0.088
Class 4	0.15	0.150	<0.001	0.067

Table A4.2

Equal Perception Model With Rater Effects, Fit Ordered Perception Model

Parameter Estimates				
Parameter	Value	Estimate	Deviation	%Deviation
d_{12}	1.00	1.016	0.016	1.570
d_{13}	2.00	2.005	0.005	0.236
d_{14}	3.00	3.019	0.019	0.632
d_{22}	2.00	1.934	-0.066	3.309
d_{23}	4.00	3.958	-0.042	1.061
d_{24}	6.00	5.968	-0.032	0.538
d_{32}	3.00	3.017	0.017	0.579
d_{33}	6.00	6.030	0.030	0.499
d_{34}	9.00	9.086	0.086	0.953
d_{42}	4.00	4.141	0.141	3.529
d_{43}	8.00	8.154	0.154	1.922
d_{44}	12.00	12.128	0.128	1.068
d_{52}	5.00	5.081	0.081	1.628
d_{53}	10.00	10.163	0.163	1.630
d_{54}	15.00	15.267	0.267	1.781
d_{62}	5.50	5.590	0.090	1.645
d_{63}	11.00	11.262	0.262	2.381
d_{64}	16.50	17.153	0.653	3.960
d_{72}	4.50	4.546	0.046	1.023
d_{73}	9.00	9.012	0.012	0.135
d_{74}	13.50	13.608	0.108	0.797
d_{82}	3.50	3.541	0.041	1.180
d_{83}	7.00	7.048	0.048	0.686

d_{84}	10.50	10.515	0.015	0.146
d_{92}	2.50	2.470	-0.030	1.184
d_{93}	5.00	4.984	-0.016	0.322
d_{94}	7.50	7.504	0.004	0.056
d_{102}	1.50	1.520	0.020	1.354
d_{103}	3.00	3.015	0.015	0.491
d_{104}	4.50	4.493	-0.007	0.165
c_{11}	0.50	0.501	0.001	0.102
c_{12}	1.50	1.509	0.009	0.609
c_{13}	2.50	2.513	0.013	0.525
c_{21}	0.00	-0.032	-0.032	-
c_{22}	2.00	1.956	-0.044	2.198
c_{23}	4.00	3.953	-0.047	1.171
c_{31}	2.50	2.522	0.022	0.890
c_{32}	5.50	5.542	0.042	0.769
c_{33}	8.50	8.564	0.064	0.748
c_{41}	3.00	3.159	0.159	5.292
c_{42}	6.00	6.165	0.165	2.744
c_{43}	9.00	9.134	0.134	1.484
c_{51}	4.50	4.579	0.079	1.761
c_{52}	9.50	9.676	0.176	1.857
c_{53}	14.50	14.773	0.273	1.883
c_{61}	0.75	0.751	0.001	0.084
c_{62}	6.25	6.351	0.101	1.615
c_{63}	11.75	12.031	0.281	2.388
c_{71}	2.25	2.289	0.039	1.716
c_{72}	6.75	6.793	0.043	0.638
c_{73}	11.25	11.280	0.030	0.271
c_{81}	0.75	0.783	0.033	4.393
c_{82}	5.25	5.292	0.042	0.793
c_{83}	9.75	9.787	0.037	0.383
c_{91}	-0.75	-0.769	-0.019	2.496
c_{92}	3.75	3.756	0.006	0.151
c_{93}	8.25	8.280	0.030	0.367
c_{101}	0.75	0.772	0.022	2.943
c_{102}	2.25	2.268	0.018	0.786
c_{103}	3.75	3.768	0.018	0.475

Latent Class Size

Parameter	Value	Estimate	Deviation	%Deviation
Class 1	0.15	0.149	-0.001	0.688
Class 2	0.35	0.348	-0.002	0.510
Class 3	0.35	0.350	<0.001	0.083
Class 4	0.15	0.153	0.003	1.686

Appendix B

Parameter Estimates, Bias/Deviation, Percent Bias/Deviation, and Mean Squared Error for the Latent Class SDT Model, BIB Design

Table B1.1

Equal Perception Model Without Rater Effects, Fit Equal Perception Model

Parameter Estimates					
Parameter	Value	Estimate	Bias	%Bias	MSE
d_1	1.00	0.959	-0.041	4.126	0.041
d_2	2.00	1.986	-0.014	0.699	0.135
d_3	3.00	2.962	-0.038	1.272	0.296
d_4	4.00	3.905	-0.095	2.387	0.606
d_5	5.00	4.740	-0.260	5.210	0.643
d_6	5.50	5.173	-0.327	5.951	0.772
d_7	4.50	4.488	-0.012	0.274	0.749
d_8	3.50	3.661	0.161	4.601	0.394
d_9	2.50	2.477	-0.023	0.914	0.162
d_{10}	1.50	1.436	-0.064	4.267	0.067
c_{11}	0.50	0.394	-0.106	21.155	0.112
c_{12}	1.50	1.415	-0.085	5.656	0.127
c_{13}	2.50	2.434	-0.066	2.625	0.150
c_{21}	1.00	0.922	-0.078	7.761	0.179
c_{22}	3.00	2.961	-0.039	1.310	0.339
c_{23}	5.00	5.028	0.028	0.555	0.629
c_{31}	1.50	1.239	-0.261	17.373	0.366
c_{32}	4.50	4.367	-0.133	2.965	0.739
c_{33}	7.50	7.522	0.022	0.293	1.663
c_{41}	2.00	1.631	-0.369	18.436	0.537
c_{42}	6.00	5.887	-0.113	1.887	1.784
c_{43}	10.00	10.094	0.094	0.939	3.758
c_{51}	2.50	1.853	-0.647	25.877	0.885
c_{52}	7.50	7.008	-0.492	6.561	1.586
c_{53}	12.50	12.345	-0.155	1.237	3.756
c_{61}	2.75	1.968	-0.782	28.419	1.291
c_{62}	8.25	7.740	-0.510	6.176	2.177
c_{63}	13.75	13.499	-0.251	1.828	4.292
c_{71}	2.25	1.880	-0.370	16.437	0.847
c_{72}	6.75	6.655	-0.095	1.414	1.861
c_{73}	11.25	11.563	0.313	2.784	4.193
c_{81}	1.75	1.640	-0.110	6.276	0.385
c_{82}	5.25	5.415	0.165	3.142	0.880
c_{83}	8.75	9.335	0.585	6.684	2.418

c_{91}	1.25	1.036	-0.214	17.141	0.298
c_{92}	3.75	3.696	-0.054	1.440	0.479
c_{93}	6.25	6.281	0.031	0.502	0.921
c_{101}	0.75	0.654	-0.096	12.774	0.125
c_{102}	2.25	2.159	-0.091	4.064	0.183
c_{103}	3.75	3.674	-0.076	2.024	0.337
		Latent Class Size			
Parameter	Value	Estimate	Bias	%Bias	MSE
Class 1	0.15	0.173	0.023	15.459	0.001
Class 2	0.35	0.332	-0.018	5.158	0.001
Class 3	0.35	0.325	-0.025	7.201	0.002
Class 4	0.15	0.170	0.020	13.379	0.001

Table B1.2

Equal Perception Model With Rater Effects, Fit Equal Perception Model

		Parameter Estimates			
Parameter	Value	Estimate	Bias	%Bias	MSE
d_1	1.00	0.958	-0.042	4.203	0.055
d_2	2.00	1.978	-0.022	1.080	0.166
d_3	3.00	2.944	-0.056	1.876	0.366
d_4	4.00	3.819	-0.181	4.534	0.535
d_5	5.00	4.111	-0.889	17.785	1.221
d_6	5.50	4.246	-1.254	22.799	2.009
d_7	4.50	4.526	0.026	0.575	0.639
d_8	3.50	3.623	0.123	3.519	0.612
d_9	2.50	2.464	-0.036	1.432	0.289
d_{10}	1.50	1.457	-0.043	2.897	0.071
c_{11}	0.50	0.413	-0.087	17.347	0.136
c_{12}	1.50	1.415	-0.085	5.655	0.152
c_{13}	2.50	2.408	-0.092	3.662	0.196
c_{21}	0.00	-0.124	-0.124	-	0.169
c_{22}	2.00	1.945	-0.055	2.750	0.319
c_{23}	4.00	3.970	-0.030	0.753	0.556
c_{31}	2.50	2.324	-0.176	7.036	0.569
c_{32}	5.50	5.432	-0.068	1.237	1.347
c_{33}	8.50	8.542	0.042	0.497	2.395
c_{41}	3.00	2.619	-0.381	12.705	0.741
c_{42}	6.00	5.583	-0.417	6.954	1.390
c_{43}	9.00	8.734	-0.266	2.960	2.474
c_{51}	4.50	3.315	-1.185	26.339	1.980
c_{52}	9.50	7.733	-1.767	18.601	4.616
c_{53}	14.50	12.066	-2.434	16.787	8.824
c_{61}	0.75	0.333	-0.417	55.561	0.492
c_{62}	6.25	4.848	-1.402	22.429	2.877

c_{63}	11.75	9.431	-2.319	19.734	7.497
c_{71}	2.25	1.789	-0.461	20.473	0.910
c_{72}	6.75	6.728	-0.022	0.328	1.999
c_{73}	11.25	11.563	0.313	2.779	4.437
c_{81}	0.75	0.539	-0.211	28.126	0.384
c_{82}	5.25	5.332	0.082	1.558	1.482
c_{83}	9.75	10.207	0.457	4.692	3.960
c_{91}	-0.75	-0.968	-0.218	29.130	0.236
c_{92}	3.75	3.680	-0.070	1.861	0.783
c_{93}	8.25	8.246	-0.004	0.046	2.081
c_{101}	0.75	0.640	-0.110	14.629	0.149
c_{102}	2.25	2.167	-0.083	3.684	0.192
c_{103}	3.75	3.682	-0.068	1.804	0.268
		Latent Class Size			
Parameter	Value	Estimate	Bias	%Bias	MSE
Class 1	0.15	0.172	0.022	14.881	0.001
Class 2	0.35	0.333	-0.017	4.825	0.002
Class 3	0.35	0.327	-0.023	6.604	0.002
Class 4	0.15	0.168	0.018	11.787	0.001

Table B2.1

Ordered Perception Model Without Rater Effects, Fit Ordered Perception Model

		Parameter Estimates			
Parameter	Value	Estimate	Bias	%Bias	MSE
d_{12}	1.00	1.110	0.110	11.033	0.790
d_{13}	2.00	1.977	-0.023	1.160	0.652
d_{14}	3.00	3.068	0.068	2.272	1.277
d_{22}	1.00	1.199	0.199	19.857	1.196
d_{23}	3.00	3.057	0.057	1.883	0.965
d_{24}	5.00	5.214	0.214	4.284	1.974
d_{32}	3.00	3.320	0.320	10.654	2.316
d_{33}	6.00	6.252	0.252	4.205	2.643
d_{34}	7.00	7.464	0.464	6.625	3.295
d_{42}	1.00	1.163	0.163	16.274	1.092
d_{43}	5.00	5.135	0.135	2.700	2.175
d_{44}	6.00	6.206	0.206	3.429	2.327
d_{52}	5.00	4.509	-0.491	9.814	2.413
d_{53}	6.00	5.087	-0.913	15.210	3.548
d_{54}	11.00	9.946	-1.054	9.586	6.623
d_{62}	5.50	4.541	-0.959	17.442	3.361
d_{63}	11.00	9.305	-1.695	15.409	7.338
d_{64}	16.50	13.905	-2.595	15.727	12.494
d_{72}	1.00	1.318	0.318	31.823	1.120
d_{73}	5.50	5.738	0.238	4.336	2.051

d_{74}	6.50	6.819	0.319	4.913	2.630
d_{82}	1.00	1.089	0.089	8.938	1.163
d_{83}	4.50	4.668	0.168	3.743	1.284
d_{84}	8.00	8.226	0.226	2.821	3.961
d_{92}	2.50	2.445	-0.055	2.216	1.768
d_{93}	3.50	3.274	-0.226	6.468	1.539
d_{94}	6.00	6.145	0.145	2.421	2.915
d_{102}	1.50	1.778	0.278	18.516	1.444
d_{103}	3.00	3.137	0.137	4.562	1.067
d_{104}	4.00	4.201	0.201	5.023	1.587
c_{11}	0.50	0.493	-0.007	1.365	0.437
c_{12}	1.50	1.549	0.049	3.269	0.487
c_{13}	2.50	2.581	0.081	3.236	0.516
c_{21}	0.50	0.466	-0.034	6.756	0.421
c_{22}	2.00	2.080	0.080	3.996	0.693
c_{23}	4.00	4.298	0.298	7.442	0.917
c_{31}	1.50	1.227	-0.273	18.233	0.978
c_{32}	4.50	4.727	0.227	5.051	2.036
c_{33}	6.50	6.913	0.413	6.350	2.656
c_{41}	0.50	0.425	-0.075	15.032	0.338
c_{42}	3.00	3.077	0.077	2.552	0.979
c_{43}	5.50	5.731	0.231	4.205	1.887
c_{51}	2.50	1.534	-0.966	38.637	2.267
c_{52}	5.50	4.802	-0.698	12.695	2.919
c_{53}	8.50	8.245	-0.255	2.997	3.383
c_{61}	2.75	1.367	-1.383	50.287	2.945
c_{62}	8.25	6.973	-1.277	15.482	5.280
c_{63}	13.75	12.404	-1.346	9.787	6.304
c_{71}	0.50	0.468	-0.032	6.462	0.394
c_{72}	3.25	3.395	0.145	4.473	1.116
c_{73}	6.00	6.286	0.286	4.769	1.960
c_{81}	0.50	0.459	-0.041	8.245	0.390
c_{82}	2.75	2.882	0.132	4.817	0.791
c_{83}	6.25	6.880	0.630	10.079	2.275
c_{91}	1.25	1.000	-0.250	19.996	1.056
c_{92}	3.00	2.855	-0.145	4.833	1.454
c_{93}	4.75	4.749	-0.001	0.028	1.679
c_{101}	0.75	0.750	<0.001	0.053	0.616
c_{102}	2.25	2.412	0.162	7.179	0.893
c_{103}	3.50	3.730	0.230	6.578	1.066

Latent Class Size

Parameter	Value	Estimate	Bias	%Bias	MSE
Class 1	0.15	0.195	0.045	30.218	0.004
Class 2	0.35	0.310	-0.040	11.570	0.004
Class 3	0.35	0.310	-0.040	11.510	0.003
Class 4	0.15	0.185	0.035	23.636	0.002

Table B2.2

Ordered Perception Model With Rater Effects, Fit Ordered Perception Model

Parameter Estimates					
Parameter	Value	Estimate	Bias	%Bias	MSE
d_{12}	1.00	1.317	0.317	31.651	1.315
d_{13}	2.00	2.121	0.121	6.055	0.896
d_{14}	3.00	3.270	0.270	9.004	1.746
d_{22}	1.00	1.228	0.228	22.845	1.619
d_{23}	3.00	3.319	0.319	10.633	1.697
d_{24}	5.00	5.207	0.207	4.141	2.074
d_{32}	3.00	2.547	-0.453	15.109	2.274
d_{33}	6.00	5.537	-0.463	7.715	2.327
d_{34}	7.00	6.741	-0.259	3.699	2.951
d_{42}	1.00	1.466	0.466	46.626	2.352
d_{43}	5.00	5.195	0.195	3.895	2.416
d_{44}	6.00	6.372	0.372	6.192	3.082
d_{52}	5.00	2.862	-2.138	42.757	7.439
d_{53}	6.00	3.577	-2.423	40.387	7.863
d_{54}	11.00	7.768	-3.232	29.386	15.039
d_{62}	5.50	4.031	-1.469	26.702	4.811
d_{63}	11.00	7.431	-3.569	32.445	16.635
d_{64}	16.50	10.990	-5.510	33.395	36.072
d_{72}	1.00	1.761	0.761	76.063	2.808
d_{73}	5.50	5.999	0.499	9.079	2.167
d_{74}	6.50	7.475	0.975	14.997	3.795
d_{82}	1.00	1.614	0.614	61.367	2.288
d_{83}	4.50	4.973	0.473	10.509	2.798
d_{84}	8.00	8.422	0.422	5.273	4.670
d_{92}	2.50	2.974	0.474	18.975	2.666
d_{93}	3.50	3.640	0.140	4.002	2.236
d_{94}	6.00	6.527	0.527	8.788	3.940
d_{102}	1.50	1.975	0.475	31.648	2.244
d_{103}	3.00	3.128	0.128	4.251	1.614
d_{104}	4.00	4.278	0.278	6.948	2.275
c_{11}	0.50	0.602	0.102	20.332	0.696
c_{12}	1.50	1.678	0.178	11.866	0.781
c_{13}	2.50	2.737	0.237	9.470	0.840
c_{21}	-0.50	-0.598	-0.098	19.554	0.365
c_{22}	1.00	1.037	0.037	3.715	0.676
c_{23}	3.00	3.302	0.302	10.076	1.082
c_{31}	2.50	1.788	-0.712	28.466	1.834
c_{32}	5.50	5.050	-0.450	8.181	2.038
c_{33}	7.50	7.313	-0.187	2.492	2.141

c_{41}	1.50	1.613	0.113	7.550	0.959
c_{42}	3.00	3.202	0.202	6.722	1.435
c_{43}	4.50	4.796	0.296	6.568	1.983
c_{51}	4.50	2.115	-2.385	52.998	7.319
c_{52}	7.50	5.428	-2.072	27.627	6.545
c_{53}	10.50	7.925	-2.575	24.527	10.117
c_{61}	0.75	0.047	-0.703	93.763	1.434
c_{62}	6.25	4.338	-1.912	30.595	6.807
c_{63}	11.75	8.668	-3.082	26.228	14.099
c_{71}	0.50	0.654	0.154	30.837	0.775
c_{72}	3.25	3.790	0.540	16.628	1.731
c_{73}	6.00	6.849	0.849	14.142	2.810
c_{81}	-0.50	-0.537	-0.037	7.479	0.359
c_{82}	2.75	3.244	0.494	17.961	1.776
c_{83}	7.25	8.141	0.891	12.290	4.029
c_{91}	-0.75	-1.089	-0.339	45.161	0.667
c_{92}	3.00	3.232	0.232	7.735	2.226
c_{93}	6.75	7.595	0.845	12.521	3.667
c_{101}	0.75	0.795	0.045	6.004	0.897
c_{102}	2.25	2.447	0.197	8.739	1.371
c_{103}	3.50	3.767	0.267	7.643	1.519

Latent Class Size

Parameter	Value	Estimate	Bias	%Bias	MSE
Class 1	0.15	0.219	0.069	46.298	0.010
Class 2	0.35	0.288	-0.062	17.695	0.008
Class 3	0.35	0.296	-0.054	15.312	0.005
Class 4	0.15	0.196	0.046	30.719	0.005

Table B3.1

Ordered Perception Model Without Rater Effects, Fit Equal Perception Model

Parameter Estimates

Parameter	Value	Estimate	Deviation	%Deviation
d_{12}	1.00	0.850	-0.150	15.014
d_{13}	2.00	1.700	-0.300	15.014
d_{14}	3.00	2.550	-0.450	15.014
d_{22}	1.00	1.546	0.546	54.590
d_{23}	3.00	3.092	0.092	3.060
d_{24}	5.00	4.638	-0.362	7.246
d_{32}	3.00	2.230	-0.770	25.663
d_{33}	6.00	4.460	-1.540	25.663
d_{34}	7.00	6.690	-0.310	4.423
d_{42}	1.00	2.208	1.208	120.831
d_{43}	5.00	4.417	-0.583	11.668
d_{44}	6.00	6.625	0.625	10.416

d_{52}	5.00	2.007	-2.993	59.867
d_{53}	6.00	4.013	-1.987	33.111
d_{54}	11.00	6.020	-4.980	45.273
d_{62}	5.50	3.972	-1.528	27.778
d_{63}	11.00	7.944	-3.056	27.778
d_{64}	16.50	11.917	-4.583	27.778
d_{72}	1.00	2.417	1.417	141.736
d_{73}	5.50	4.835	-0.665	12.096
d_{74}	6.50	7.252	0.752	11.571
d_{82}	1.00	2.548	1.548	154.835
d_{83}	4.50	5.097	0.597	13.260
d_{84}	8.00	7.645	-0.355	4.437
d_{92}	2.50	1.439	-1.061	42.424
d_{93}	3.50	2.879	-0.621	17.748
d_{94}	6.00	4.318	-1.682	28.030
d_{102}	1.50	1.221	-0.279	18.619
d_{103}	3.00	2.441	-0.559	18.619
d_{104}	4.00	3.662	-0.338	8.447
c_{11}	0.50	0.250	-0.250	49.913
c_{12}	1.50	1.270	-0.230	15.334
c_{13}	2.50	2.262	-0.238	9.520
c_{21}	0.50	0.560	0.060	12.051
c_{22}	2.00	2.130	0.130	6.523
c_{23}	4.00	4.162	0.162	4.052
c_{31}	1.50	0.473	-1.027	68.434
c_{32}	4.50	3.394	-1.106	24.576
c_{33}	6.50	5.574	-0.926	14.240
c_{41}	0.50	0.791	0.291	58.214
c_{42}	3.00	3.204	0.204	6.803
c_{43}	5.50	5.598	0.098	1.789
c_{51}	2.50	0.224	-2.276	91.041
c_{52}	5.50	2.878	-2.622	47.675
c_{53}	8.50	5.619	-2.881	33.897
c_{61}	2.75	0.693	-2.057	74.785
c_{62}	8.25	5.852	-2.398	29.067
c_{63}	13.75	10.859	-2.891	21.023
c_{71}	0.50	0.778	0.278	55.575
c_{72}	3.25	3.464	0.214	6.596
c_{73}	6.00	6.083	0.083	1.379
c_{81}	0.50	1.061	0.561	112.207
c_{82}	2.75	3.518	0.768	27.913
c_{83}	6.25	6.924	0.674	10.779
c_{91}	1.25	0.326	-0.924	73.956
c_{92}	3.00	2.065	-0.935	31.162
c_{93}	4.75	3.803	-0.947	19.938
c_{101}	0.75	0.302	-0.448	59.784

c_{102}	2.25	1.837	-0.413	18.334
c_{103}	3.50	3.125	-0.375	10.712
Latent Class Size				
Parameter	Value	Estimate	Deviation	%Deviation
Class 1	0.15	0.248	0.098	65.250
Class 2	0.35	0.268	-0.082	23.534
Class 3	0.35	0.268	-0.082	23.434
Class 4	0.15	0.217	0.067	44.341

Table B3.2

Ordered Perception Model With Rater Effects, Fit Equal Perception Model

Parameter Estimates				
Parameter	Value	Estimate	Deviation	%Deviation
d_{12}	1.00	0.797	-0.203	20.258
d_{13}	2.00	1.595	-0.405	20.258
d_{14}	3.00	2.392	-0.608	20.258
d_{22}	1.00	1.436	0.436	43.567
d_{23}	3.00	2.871	-0.129	4.288
d_{24}	5.00	4.307	-0.693	13.860
d_{32}	3.00	1.913	-1.087	36.238
d_{33}	6.00	3.826	-2.174	36.238
d_{34}	7.00	5.739	-1.261	18.020
d_{42}	1.00	2.047	1.047	104.733
d_{43}	5.00	4.095	-0.905	18.107
d_{44}	6.00	6.142	0.142	2.367
d_{52}	5.00	1.570	-3.430	68.605
d_{53}	6.00	3.139	-2.861	47.676
d_{54}	11.00	4.709	-6.291	57.189
d_{62}	5.50	2.920	-2.580	46.906
d_{63}	11.00	5.840	-5.160	46.906
d_{64}	16.50	8.761	-7.739	46.906
d_{72}	1.00	2.319	1.319	131.937
d_{73}	5.50	4.639	-0.861	15.659
d_{74}	6.50	6.958	0.458	7.048
d_{82}	1.00	2.321	1.321	132.071
d_{83}	4.50	4.641	0.141	3.143
d_{84}	8.00	6.962	-1.038	12.973
d_{92}	2.50	1.229	-1.271	50.858
d_{93}	3.50	2.457	-1.043	29.797
d_{94}	6.00	3.686	-2.314	38.573
d_{102}	1.50	1.081	-0.419	27.953
d_{103}	3.00	2.161	-0.839	27.953
d_{104}	4.00	3.242	-0.758	18.947
c_{11}	0.50	0.163	-0.337	67.358

c_{12}	1.50	1.187	-0.313	20.849
c_{13}	2.50	2.193	-0.307	12.281
c_{21}	-0.50	-0.651	-0.151	30.170
c_{22}	1.00	0.904	-0.096	9.594
c_{23}	3.00	3.022	0.022	0.731
c_{31}	2.50	0.985	-1.515	60.583
c_{32}	5.50	3.930	-1.570	28.546
c_{33}	7.50	6.205	-1.295	17.269
c_{41}	1.50	1.581	0.081	5.370
c_{42}	3.00	2.972	-0.028	0.922
c_{43}	4.50	4.395	-0.105	2.331
c_{51}	4.50	1.094	-3.406	75.693
c_{52}	7.50	4.197	-3.303	44.037
c_{53}	10.50	5.670	-4.830	46.000
c_{61}	0.75	-0.555	-1.305	173.978
c_{62}	6.25	2.892	-3.358	53.733
c_{63}	11.75	7.312	-4.438	37.766
c_{71}	0.50	0.582	0.082	16.325
c_{72}	3.25	3.385	0.135	4.163
c_{73}	6.00	6.193	0.193	3.212
c_{81}	-0.50	-0.540	-0.040	8.036
c_{82}	2.75	3.254	0.504	18.309
c_{83}	7.25	7.300	0.050	0.695
c_{91}	-0.75	-1.627	-0.877	116.936
c_{92}	3.00	1.762	-1.238	41.263
c_{93}	6.75	5.244	-1.506	22.315
c_{101}	0.75	0.145	-0.605	80.708
c_{102}	2.25	1.619	-0.631	28.031
c_{103}	3.50	2.892	-0.608	17.376

Latent Class Size

Parameter	Value	Estimate	Deviation	%Deviation
Class 1	0.15	0.302	0.152	101.039
Class 2	0.35	0.205	-0.145	41.305
Class 3	0.35	0.226	-0.124	35.343
Class 4	0.15	0.267	0.117	77.807

Table B4.1

Equal Perception Model Without Rater Effects, Fit Ordered Perception Model

Parameter Estimates				
Parameter	Value	Estimate	Deviation	%Deviation
d_{12}	1.00	1.061	0.061	6.116
d_{13}	2.00	2.026	0.026	1.276
d_{14}	3.00	3.032	0.032	1.074
d_{22}	2.00	2.237	0.237	11.827
d_{23}	4.00	4.169	0.169	4.227
d_{24}	6.00	6.306	0.306	5.107
d_{32}	3.00	3.128	0.128	4.259
d_{33}	6.00	6.120	0.120	1.994
d_{34}	9.00	9.299	0.299	3.324
d_{42}	4.00	4.213	0.213	5.319
d_{43}	8.00	8.112	0.112	1.397
d_{44}	12.00	12.255	0.255	2.123
d_{52}	5.00	5.150	0.150	2.990
d_{53}	10.00	9.743	-0.257	2.569
d_{54}	15.00	14.594	-0.406	2.704
d_{62}	5.50	5.337	-0.163	2.958
d_{63}	11.00	10.376	-0.624	5.675
d_{64}	16.50	15.724	-0.776	4.701
d_{72}	4.50	4.706	0.206	4.567
d_{73}	9.00	9.177	0.177	1.972
d_{74}	13.50	13.710	0.210	1.552
d_{82}	3.50	3.670	0.170	4.848
d_{83}	7.00	7.480	0.480	6.860
d_{84}	10.50	11.451	0.951	9.059
d_{92}	2.50	2.421	-0.079	3.165
d_{93}	5.00	5.032	0.032	0.642
d_{94}	7.50	7.691	0.191	2.544
d_{102}	1.50	1.396	-0.104	6.920
d_{103}	3.00	2.855	-0.145	4.825
d_{104}	4.50	4.449	-0.051	1.127
c_{11}	0.50	0.463	-0.037	7.398
c_{12}	1.50	1.517	0.017	1.140
c_{13}	2.50	2.564	0.064	2.555
c_{21}	1.00	1.075	0.075	7.481
c_{22}	3.00	3.174	0.174	5.795
c_{23}	5.00	5.311	0.311	6.210
c_{31}	1.50	1.301	-0.199	13.261
c_{32}	4.50	4.542	0.042	0.935
c_{33}	7.50	7.847	0.347	4.623
c_{41}	2.00	1.781	-0.219	10.969
c_{42}	6.00	6.181	0.181	3.009

c_{43}	10.00	10.541	0.541	5.409
c_{51}	2.50	1.941	-0.559	22.371
c_{52}	7.50	7.381	-0.119	1.584
c_{53}	12.50	12.765	0.265	2.117
c_{61}	2.75	1.986	-0.764	27.769
c_{62}	8.25	7.829	-0.421	5.097
c_{63}	13.75	13.710	-0.040	0.289
c_{71}	2.25	1.867	-0.383	17.043
c_{72}	6.75	6.877	0.127	1.888
c_{73}	11.25	11.869	0.619	5.502
c_{81}	1.75	1.627	-0.123	7.012
c_{82}	5.25	5.508	0.258	4.923
c_{83}	8.75	9.672	0.922	10.537
c_{91}	1.25	0.993	-0.257	20.549
c_{92}	3.75	3.720	-0.030	0.800
c_{93}	6.25	6.423	0.173	2.771
c_{101}	0.75	0.614	-0.136	18.183
c_{102}	2.25	2.156	-0.094	4.175
c_{103}	3.75	3.729	-0.021	0.548
Latent Class Size				
Parameter	Value	Estimate	Deviation	%Deviation
Class 1	0.15	0.175	0.025	16.853
Class 2	0.35	0.329	-0.021	6.082
Class 3	0.35	0.324	-0.026	7.320
Class 4	0.15	0.172	0.022	14.420

Table B4.2

Equal Perception Model With Rater Effects, Fit Ordered Perception Model

Parameter Estimates				
Parameter	Value	Estimate	Deviation	%Deviation
d_{12}	1.00	1.036	0.036	3.582
d_{13}	2.00	1.932	-0.068	3.420
d_{14}	3.00	3.048	0.048	1.614
d_{22}	2.00	2.043	0.043	2.146
d_{23}	4.00	4.136	0.136	3.389
d_{24}	6.00	6.283	0.283	4.723
d_{32}	3.00	2.871	-0.129	4.311
d_{33}	6.00	5.852	-0.148	2.468
d_{34}	9.00	9.323	0.323	3.587
d_{42}	4.00	3.715	-0.285	7.113
d_{43}	8.00	7.744	-0.256	3.196
d_{44}	12.00	11.441	-0.559	4.661
d_{52}	5.00	3.787	-1.213	24.269
d_{53}	10.00	7.641	-2.359	23.585

d_{54}	15.00	12.141	-2.859	19.057
d_{62}	5.50	4.839	-0.661	12.023
d_{63}	11.00	8.876	-2.124	19.310
d_{64}	16.50	12.882	-3.618	21.926
d_{72}	4.50	4.746	0.246	5.461
d_{73}	9.00	9.431	0.431	4.790
d_{74}	13.50	14.132	0.632	4.680
d_{82}	3.50	4.268	0.768	21.954
d_{83}	7.00	7.779	0.779	11.123
d_{84}	10.50	11.873	1.373	13.077
d_{92}	2.50	2.941	0.441	17.655
d_{93}	5.00	5.424	0.424	8.477
d_{94}	7.50	8.091	0.591	7.877
d_{102}	1.50	1.639	0.139	9.233
d_{103}	3.00	3.083	0.083	2.770
d_{104}	4.50	4.683	0.183	4.063
c_{11}	0.50	0.450	-0.050	10.030
c_{12}	1.50	1.478	-0.022	1.488
c_{13}	2.50	2.500	<0.001	0.005
c_{21}	0.00	-0.168	-0.168	-
c_{22}	2.00	2.013	0.013	0.672
c_{23}	4.00	4.152	0.152	3.808
c_{31}	2.50	2.245	-0.255	10.189
c_{32}	5.50	5.476	-0.024	0.436
c_{33}	8.50	8.998	0.498	5.860
c_{41}	3.00	2.516	-0.484	16.149
c_{42}	6.00	5.597	-0.403	6.708
c_{43}	9.00	8.894	-0.106	1.173
c_{51}	4.50	2.970	-1.530	34.005
c_{52}	9.50	7.254	-2.246	23.644
c_{53}	14.50	11.922	-2.578	17.777
c_{61}	0.75	0.365	-0.385	51.361
c_{62}	6.25	5.406	-0.844	13.499
c_{63}	11.75	9.918	-1.832	15.591
c_{71}	2.25	1.806	-0.444	19.728
c_{72}	6.75	7.042	0.292	4.325
c_{73}	11.25	12.158	0.908	8.070
c_{81}	0.75	0.690	-0.060	7.993
c_{82}	5.25	5.966	0.716	13.632
c_{83}	9.75	11.214	1.464	15.012
c_{91}	-0.75	-0.987	-0.237	31.652
c_{92}	3.75	4.168	0.418	11.144
c_{93}	8.25	9.042	0.792	9.599
c_{101}	0.75	0.755	0.005	0.723
c_{102}	2.25	2.343	0.093	4.132
c_{103}	3.75	3.915	0.165	4.408

Latent Class Size				
Parameter	Value	Estimate	Deviation	%Deviation
Class 1	0.15	0.177	0.027	17.920
Class 2	0.35	0.325	-0.025	7.088
Class 3	0.35	0.322	-0.028	7.880
Class 4	0.15	0.176	0.026	17.007

Appendix C

Parameter Estimates and Standard Error for the Ordered Perception Latent Class SDT Model, Real Data

Table C1.

Results for Ordered Perception SDT Model, One Essay, Language Test

Parameter Estimates				
Parameter	Estimate	SE	*z*-value	*p*-value
d_{12}	6.412	2.679	2.394	0.017
d_{13}	9.438	2.760	3.420	0.001
d_{14}	12.513	2.897	4.320	0.000
d_{15}	17.272	3.648	4.735	0.000
d_{22}	4.109	2.162	1.901	**0.057**
d_{23}	8.507	2.734	3.112	0.002
d_{23}	14.077	3.324	4.235	0.000
d_{25}	15.613	3.427	4.555	0.000
d_{32}	3.957	1.241	3.188	0.001
d_{33}	8.407	2.054	4.093	0.000
d_{34}	10.361	2.108	4.914	0.000
d_{35}	13.739	2.376	5.782	0.000
d_{42}	3.241	0.967	3.353	0.001
d_{43}	7.831	1.696	4.618	0.000
d_{44}	11.332	1.910	5.932	0.000
d_{45}	13.141	1.980	6.636	0.000
d_{52}	5.296	1.478	3.583	0.000
d_{53}	8.126	1.748	4.649	0.000
d_{54}	11.754	2.025	5.804	0.000
d_{55}	16.755	2.874	5.830	0.000
d_{62}	6.608	2.540	2.602	0.009
d_{63}	10.466	2.712	3.859	0.000
d_{64}	15.439	3.237	4.770	0.000
d_{65}	19.150	3.408	5.620	0.000
d_{72}	4.181	1.438	2.907	0.004
d_{73}	7.082	1.528	4.634	0.000
d_{74}	9.114	1.565	5.825	0.000
d_{75}	12.877	1.816	7.092	0.000
d_{82}	3.825	1.203	3.180	0.002
d_{83}	4.701	1.143	4.111	0.000
d_{84}	8.768	2.071	4.233	0.000
d_{85}	13.080	3.433	3.810	0.000
d_{92}	5.825	2.027	2.874	0.004
d_{93}	10.344	2.685	3.852	0.000

d_{94}	17.815	3.466	5.140	0.000
d_{95}	20.210	3.558	5.680	0.000
d_{102}	6.144	2.411	2.548	0.011
d_{103}	9.383	2.782	3.373	0.001
d_{104}	13.477	3.257	4.137	0.000
d_{105}	18.779	3.946	4.759	0.000
d_{112}	4.986	1.426	3.497	0.000
d_{113}	8.999	1.863	4.829	0.000
d_{114}	12.937	2.184	5.924	0.000
d_{115}	15.721	2.190	7.179	0.000
d_{122}	2.686	0.925	2.903	0.004
d_{123}	5.096	1.101	4.630	0.000
d_{124}	10.302	1.910	5.394	0.000
d_{125}	13.715	2.329	5.889	0.000
d_{132}	4.746	1.077	4.404	0.000
d_{133}	8.106	1.368	5.927	0.000
d_{134}	11.600	1.528	7.590	0.000
d_{135}	15.505	1.738	8.920	0.000
d_{142}	7.522	3.043	2.472	0.013
d_{143}	10.543	3.232	3.262	0.001
d_{144}	14.296	3.363	4.251	0.000
d_{145}	19.586	4.112	4.762	0.000
d_{152}	3.687	1.274	2.894	0.004
d_{153}	10.499	2.643	3.973	0.000
d_{154}	14.423	2.841	5.077	0.000
d_{155}	20.336	3.779	5.382	0.000
d_{162}	7.106	2.172	3.272	0.001
d_{163}	10.829	2.358	4.592	0.000
d_{164}	15.234	2.602	5.854	0.000
d_{165}	21.232	3.434	6.183	0.000
d_{172}	4.848	1.792	2.706	0.007
d_{173}	10.997	2.508	4.385	0.000
d_{174}	14.656	2.719	5.390	0.000
d_{175}	18.265	3.098	5.895	0.000
d_{182}	5.220	2.471	2.112	0.035
d_{183}	7.289	2.556	2.852	0.004
d_{184}	10.174	2.634	3.862	0.000
d_{185}	15.691	3.648	4.301	0.000
d_{192}	5.348	2.149	2.489	0.013
d_{193}	9.600	2.723	3.525	0.000
d_{194}	13.651	3.169	4.308	0.000
d_{195}	15.745	3.234	4.869	0.000
d_{202}	5.981	2.179	2.745	0.006
d_{203}	9.709	2.584	3.757	0.000
d_{204}	12.981	2.905	4.468	0.000
d_{205}	20.230	3.927	5.152	0.000

d_{212}	4.556	2.057	2.215	0.027
d_{213}	9.010	2.532	3.558	0.000
d_{214}	13.524	2.947	4.590	0.000
d_{215}	19.865	4.188	4.744	0.000
d_{222}	4.451	2.540	1.752	**0.080**
d_{223}	7.602	2.882	2.638	0.008
d_{224}	14.074	3.634	3.873	0.000
d_{225}	15.998	3.716	4.305	0.000
d_{232}	5.194	2.733	1.901	**0.057**
d_{233}	7.868	2.812	2.798	0.005
d_{233}	9.305	2.872	3.240	0.001
d_{235}	11.050	2.891	3.822	0.000
d_{242}	4.500	2.442	1.843	**0.065**
d_{243}	9.845	3.107	3.169	0.002
d_{244}	13.291	3.369	3.945	0.000
d_{245}	16.618	3.529	4.710	0.000
d_{252}	5.510	2.285	2.411	0.016
d_{253}	11.729	3.273	3.584	0.000
d_{253}	14.384	3.416	4.211	0.000
d_{254}	17.500	3.598	4.864	0.000
d_{262}	2.424	1.568	1.546	**0.120**
d_{263}	3.970	1.571	2.527	0.012
d_{264}	7.543	2.318	3.254	0.001
d_{265}	9.002	2.293	3.926	0.000
d_{272}	4.630	2.736	1.693	**0.091**
d_{273}	9.142	3.207	2.851	0.004
d_{274}	11.948	3.346	3.571	0.000
d_{275}	14.155	3.518	4.024	0.000
c_{11}	4.425	2.447	-1.809	**0.071**
c_{12}	8.032	2.694	-2.981	0.003
c_{13}	10.873	2.777	-3.916	0.000
c_{14}	13.730	2.892	-4.747	0.000
c_{21}	0.419	0.573	-0.731	**0.460**
c_{22}	4.496	2.127	-2.114	0.035
c_{23}	9.885	2.814	-3.513	0.000
c_{24}	14.865	3.362	-4.422	0.000
c_{31}	1.312	0.587	-2.233	0.026
c_{32}	4.397	1.198	-3.670	0.000
c_{33}	9.217	2.012	-4.581	0.000
c_{34}	13.236	2.248	-5.887	0.000
c_{41}	1.729	0.646	-2.678	0.007
c_{42}	5.793	1.403	-4.128	0.000
c_{43}	9.865	1.803	-5.472	0.000
c_{44}	12.695	1.914	-6.634	0.000
c_{51}	2.373	1.014	-2.339	0.019
c_{52}	6.980	1.605	-4.349	0.000

c_{53}	10.904	1.911	-5.706	0.000
c_{54}	16.256	2.816	-5.774	0.000
c_{61}	4.904	2.443	-2.007	0.045
c_{62}	8.501	2.615	-3.250	0.001
c_{63}	14.787	3.173	-4.660	0.000
c_{64}	18.342	3.349	-5.477	0.000
c_{71}	1.422	0.462	-3.081	0.002
c_{72}	5.150	1.306	-3.942	0.000
c_{73}	8.370	1.471	-5.690	0.000
c_{74}	11.177	1.615	-6.919	0.000
c_{81}	1.615	0.729	-2.215	0.027
c_{82}	4.801	1.039	-4.619	0.000
c_{83}	9.002	1.934	-4.655	0.000
c_{84}	12.112	2.938	-4.123	0.000
c_{91}	2.315	0.789	-2.934	0.003
c_{92}	6.888	2.049	-3.362	0.001
c_{93}	13.588	2.996	-4.535	0.000
c_{94}	19.372	3.517	-5.508	0.000
c_{101}	3.650	2.233	-1.634	0.100
c_{102}	8.684	2.675	-3.247	0.001
c_{103}	14.018	3.192	-4.392	0.000
c_{104}	19.817	3.939	-5.031	0.000
c_{111}	2.796	1.141	-2.450	0.014
c_{112}	6.403	1.539	-4.161	0.000
c_{113}	10.083	1.849	-5.453	0.000
c_{114}	13.746	2.099	-6.547	0.000
c_{121}	0.349	0.417	-0.838	0.400
c_{122}	3.859	0.876	-4.404	0.000
c_{123}	8.616	1.726	-4.992	0.000
c_{124}	13.162	2.222	-5.923	0.000
c_{131}	2.262	0.710	-3.186	0.001
c_{132}	6.411	1.184	-5.414	0.000
c_{133}	9.793	1.379	-7.100	0.000
c_{134}	13.996	1.604	-8.723	0.000
c_{141}	4.418	2.446	-1.806	0.071
c_{142}	8.390	3.030	-2.769	0.006
c_{143}	12.181	3.231	-3.770	0.000
c_{144}	16.863	3.572	-4.721	0.000
c_{151}	1.824	0.832	-2.193	0.028
c_{152}	7.361	2.041	-3.606	0.000
c_{153}	12.665	2.688	-4.712	0.000
c_{154}	17.926	3.336	-5.373	0.000
c_{161}	2.161	0.763	-2.831	0.005
c_{162}	8.456	2.239	-3.776	0.000
c_{163}	12.890	2.443	-5.277	0.000
c_{164}	19.967	3.328	-5.999	0.000

c_{171}	1.364	0.537	-2.540	0.011
c_{172}	6.797	1.830	-3.713	0.000
c_{173}	12.750	2.554	-4.992	0.000
c_{174}	17.707	3.008	-5.886	0.000
c_{181}	1.897	1.077	-1.761	**0.078**
c_{182}	6.117	2.435	-2.513	0.012
c_{183}	9.029	2.537	-3.558	0.000
c_{184}	13.136	3.013	-4.360	0.000
c_{191}	0.657	0.506	-1.297	**0.190**
c_{192}	6.060	2.168	-2.795	0.005
c_{193}	9.703	2.680	-3.620	0.000
c_{194}	14.062	3.150	-4.464	0.000
c_{201}	1.688	0.768	-2.196	0.028
c_{202}	6.964	2.230	-3.123	0.002
c_{203}	11.523	2.655	-4.341	0.000
c_{204}	15.384	2.952	-5.212	0.000
c_{211}	0.913	0.512	-1.783	**0.075**
c_{212}	6.830	2.183	-3.129	0.002
c_{213}	11.453	2.686	-4.263	0.000
c_{214}	16.249	3.192	-5.090	0.000
c_{221}	3.372	2.417	-1.395	**0.160**
c_{222}	5.943	2.648	-2.244	0.025
c_{223}	11.374	3.254	-3.495	0.000
c_{224}	15.763	3.604	-4.374	0.000
c_{231}	3.738	2.468	-1.515	**0.130**
c_{232}	5.947	2.644	-2.249	0.025
c_{233}	7.763	2.720	-2.854	0.004
c_{234}	10.132	2.790	-3.632	0.000
c_{241}	-0.129	0.792	0.163	**0.870**
c_{242}	5.419	2.491	-2.176	0.030
c_{243}	11.588	3.182	-3.642	0.000
c_{244}	15.266	3.409	-4.478	0.000
c_{251}	3.444	2.035	-1.692	**0.091**
c_{252}	8.377	2.780	-3.013	0.003
c_{253}	13.028	3.286	-3.964	0.000
c_{254}	17.602	3.531	-4.985	0.000
c_{261}	0.880	1.077	-0.817	**0.410**
c_{262}	2.301	1.340	-1.718	**0.086**
c_{263}	4.538	1.490	-3.046	0.002
c_{264}	9.003	2.173	-4.143	0.000
c_{271}	3.742	2.511	-1.490	**0.140**
c_{272}	7.936	3.096	-2.563	0.010
c_{273}	11.178	3.250	-3.440	0.001
c_{274}	13.864	3.377	-4.106	0.000